GENERALIZED LINEAR LONGITUDINAL SEMI-PARAMETRIC MODELS WITH TIME DEPENDENT COVARIATES

VINEETHA WARRIYAR KODALORE VIJAYAN

Generalized Linear Longitudinal Semi-parametric Models with Time Dependent Covariates

by

© **Vineetha Warriyar K. V**

A thesis submitted to the

School of Graduate Studies

in partial fulfilment of the

requirements for the degree of

Doctor of Philosophy

Department of Mathematics and Statistics

Memorial University of Newfoundland

November 2012

St. John's Newfoundland

Abstract

Longitudinal data analysis is challenging because of the difficulties in modelling the correlations among the repeated responses, especially when the associated covariates are time dependent. Recent studies have examined correlations for both linear and discrete unbalanced longitudinal data, which are modelled following a Gaussian-type auto-regressive moving average (ARMA) class of auto-correlations. However, these studies were confined to a regression setup where the regression function is completely specified. In this thesis, we consider a semi-parametric regression setup in which the regression function involves a specified as well as an unspecified function over time. Under the ARMA type correlation structure, we provide a semi-parametric generalized quasi-likelihood (SGQL) approach for the estimation of the main regression parameters. The proposed inference approach is compared with some existing generalized estimating equation (GEE) approaches mainly through simulation studies. The linear longitudinal semi-parametric model, for its foundational nature, is discussed in detail. Theoretical details on semi-parametric estimation for longitudinal count and binary data are also provided.

Acknowledgements

I would like to express sincere thanks to my supervisor, Dr. Brajendra Sutradhar. His support, vast knowledge and logical way of thinking have been invaluable throughout my work.

Thanks to all my friends, colleagues and house mates for putting up with me. I owe a great deal for their encouragement and moral support.

A special thanks to the Department of Mathematics and Statistics at Memorial University for providing financial and academic assistance for my research, especially all administrative staff members for the help they offered me during my study.

I would like to thank all the examiners of my thesis Dr. Mary Thompson, Dr. Gary Sneddon and Dr. Alwell Oyet for their invaluable comments and suggestions.

I am so indebted to my parents, my sister, my in-laws, and my uncle Dr. Asokan. M. Variyath, who motivated me to explore the wonderful world of Statistics, for their constant support and encouragement. Indeed, I am grateful to my husband, Ranjith, for his encouragement, confidence and forbearance. Without my family's understanding and love, I would not have been able to finish this thesis. Above all, I thank Almighty God, without whose blessing I would never have been able to complete this work.

— Vineetha Warriyar K. V

I lovingly dedicate this thesis

to my husband Ranjith, who supported me every step of the way.

Contents

List of Tables

List of Figures

Chapter 1

Background of the Problem

Longitudinal studies are common in many scientific research areas such as clinical trials, economics, public health, agriculture, and so on. In these studies, the responses along with the covariates are collected from individuals over a period of time. In many cases, the time points are equally spaced. For example, (1) the Ohio asthma data [Zeger, Liang and Albert (1988)] collected from 537 children every year over a period of four years; (2) the health care utilization data [Sutradhar (2003, page 391)] collected by the General Hospital of the city of St. John's, Newfoundland, Canada, which contains the number of yearly visits to a physician by individuals over four consecutive years; and (3) the survey of labour and income dynamics (SLID) data on unemployment status among others collected by Statistics Canada [Sutradhar (2011)] every year over a period of six or more years. There are other situations where a respondent reports a response whenever an event occurs, where time points may not be equi-spaced. Because the repeated data are likely to be correlated, it is important to take such correlations into account for efficient inferences of the regression

effects involved in the model. However, the modelling of the correlations especially when the responses are discrete is difficult even if the responses are collected over equi-spaced time points. In a fixed regression setup, Sutradhar (2010) suggested a Gaussian-type ARMA class of auto-correlation models appropriate for both linear and discrete longitudinal data. These regression models however, may be inadequate in situations where a specified (or fixed) regression function may not be sufficient to interpret the responses completely. In such cases, one may extend these models by adding an unspecified non-parametric function in time with the fixed regression function. This leads to a semi-parametric regression model setup where longitudinal responses still follow a suitable correlation structure. There exists generalized estimating equation (GEE) based approaches to deal with the inferences for the aforementioned semi-parametric models in the longitudinal setup, where the modelling of longitudinal correlations are not done. In this thesis, however, we concentrate on the semi-parametric inferences for repeated data which follow a ARMA-type class of auto-correlations. In order to give a background for this semi-parametric modelling and inference problem in the longitudinal setup, we first provide the notations and an overview for the semi-parametric problem in independence setup in Sections 1.1 and 1.2. A brief overview of the same semi-parametric problem in longitudinal setup is provided in Sections 1.3 and 1.4.

1.1 Generalized linear models (GLMs)

Consider a GLM regression set up [Nelder and Wedderburn (1972)] in which an exponential family based independent responses $\{y_i\}$, $i = 1, \ldots, K$ are observed.

Let $x_i = (x_{i1}, \ldots, x_{ip})'$ be a multidimensional covariate vector corresponding to y_i for the i^{th} individual. Suppose that the mean response $\mu_i(\beta) = E(Y_i)$ is influenced by a specified fixed regression function (linear predictor) $x_i'\beta$ with $\beta = (\beta_1, \ldots, \beta_p)'$. The density of the exponential family based response y_i can be written as

$$f(y_i|\theta_i) = \exp[y_i\theta_i - a(\theta_i) + b(y_i)] \tag{1.1}$$

where $a(.)$ and $b(.)$ are known functional forms such that $b(.)$ depends only on y_i, and the canonical parameter θ_i is defined with a suitable link function $h(.)$ as

$$\theta_i = h(x_i'\beta). \tag{1.2}$$

The parameter θ_i is related to the mean response through

$$\mu_i(\beta) = E(Y_i|x_i) = a'(\theta_i) \tag{1.3}$$

where $a'(.)$ is the first derivative of $a(.)$ with respect to θ_i. Also, it follows that the variance of y_i is

$$\sigma_{ii}(\beta) = var(Y_i|x_i) = a''(\theta_i) \tag{1.4}$$

where $a''(.)$ is the second derivative of $a(.)$ with respect to θ_i.

1.1.1 Quasi-likelihood estimation for β

In the above exponential setup, the regression parameter β is involved in $\mu_i(\beta) = a'(\theta_i)$ as well as in $\sigma_{ii}(\beta) = a''(\theta_i)$. Since $\sigma_{ii}(\beta)$ is a function of the mean response, it is sufficient to estimate β involved in $\mu_i(\beta)$. When the density function is not known, and the mean and variance are given, Wedderburn (1974) proposed the quasi-likelihood

(QL) estimation approach to estimate the regression parameter. In this approach, one solves the QL estimating equation

$$\sum_{i=1}^{K} \frac{\partial a'(\theta_i)}{\partial \beta}[a''(\theta_i)]^{-1}(y_i - a'(\theta_i)) = \sum_{i=1}^{K} \frac{\partial \mu_i(\beta)}{\partial \beta}[\sigma_{ii}(\beta)]^{-1}(y_i - \mu_i(\beta)) = 0 \qquad (1.5)$$

[see also McCullagh (1983), McCullagh and Nelder (1989)]. The estimate $\hat{\beta}_{QL}$ obtained by solving (1.5) is consistent and highly efficient. This is because under the exponential family setup, the QL estimate turns out to be the likelihood estimate, which is known to be optimal (highly efficient). [Sutradhar (2010a)].

1.2 Semi-parametric GLMs

In semi-parametric models, the mean response $\mu_i(\beta)$ depends not only on a fixed regression function, but also on an unspecified (non-parametric) smooth function, namely $\gamma(z_i)$, where z_i is an auxiliary covariate which influences the response y_i. Then $\mu_i(\beta)$ becomes a function of an unknown parameter vector β and an unknown smooth function $\gamma(z_i)$, which we abbreviate as

$$\mu_i(\beta, \gamma(z_i)) = E(Y_i|x_i, z_i). \qquad (1.6)$$

In this set up, the canonical parameter θ_i defined in (1.2) has the form

$$\theta_i = h(x_i'\beta + \gamma(z_i)). \qquad (1.7)$$

It is clear that the main regression parameter β can no longer be estimated unbiasedly by ignoring the estimation of $\gamma(z_i)$. The semi-parametric GLMs are more flexible than the parametric GLMs especially when the regression function in fixed covariates is insufficient to understand the mean response.

Even though the estimation of both fixed regression parameter vector β and the non-parametric function $\gamma(.)$ are of interest, many early works [Staniswalis (1989), and Muller (1988)] concentrated on the estimation of the non-parametric mean function, which is the same as substituting $\beta = 0$ in (1.7). To deal with this type of non-parametric regression estimation there exists many kernel methods and its variants, such as the Nadaraya-Watson kernel regression estimation [Nadaraya (1964), Watson (1964), Bierens (1987), Andrews (1995)], local linear and polynomial regression [Cleveland (1979), Fan (1992, 1993), Stone (1980, 1982)], recursive kernel estimation [see e.g., Ahmad and Lin (1976), Greblicki and Krzyzak (1980)], spline smoothing [Whittaker(1923), Eubank (1988), Wahba (1990)], and nearest neighbour estimation [Royall (1966), Stone(1977)]. Among these techniques, the simpler Nadaraya-Watson kernel estimator or the local constant estimator for $\gamma(z)$ at a given covariate level $z = z_0$ involved in the linear model,

$$y_i = \gamma(z_i) + \epsilon_i, i = 1, \ldots, K, \text{ and } \epsilon_i \sim (0, \sigma_\epsilon^2)$$

has the form

$$\hat{\gamma}(z_0) = \frac{\sum_{i=1}^{K} y_i \; K^*(\frac{z_0 - z_i}{b})}{\sum_{i=1}^{K} K^*(\frac{z_0 - z_i}{b})}$$

where $K^*(.)$ is a suitable kernel density function and b is known as the bandwidth.

The selection of an appropriate bandwidth parameter b is always a problem in non-parametric regression [Silverman(1986)]. In practice, we try to use a possible value of b for which the bias and variance of the estimator will be minimum. Many data-based methods such as cross validation [see Stone (1974), Picard and Cook (1984), Ansley, Kohn, and Tharm (1991)], generalized cross validation [Craven and Wahba

(1979)] were discussed in the literature for choosing an appropriate b. Altman (1990) suggested that these commonly used bandwidth selection techniques do not perform well when the errors are correlated. Hence we excluded these techniques and followed Pagan and Ullah (1999) who proposed an optimum value for bandwidth, which minimizes the approximate mean integrated squared error. The authors recommended $b \propto n^{-1/5}$, and suggested that this value of bandwidth is the only value of b for which the bias and variance are of the same order of magnitude. Thus, as a practical choice, we will consider $b = K^{-1/5}$.

In the independence set up, the estimation of both β and $\gamma(.)$ are also extensively studied in the literature [e.g., Severini and Staniswalis (1994), Carota and Parmigiani (2002)]. Under the exponential family, for example, Severini and Staniswalis (1994) suggested a semi-parametric QL (SQL) approach for the estimation of β and $\gamma(.)$. The authors illustrated their estimation methodology using examples with linear, gamma and binary data. Note that we do not deal with (continuous) gamma data in the thesis, instead, we concentrate on modelling and inferences for linear and discrete data such as count and binary data in semi-parametric set up for independent and longitudinal responses. For convenience, we now provide semi-parametric QL estimation in details for linear, count and binary data in the independence set up.

1.2.1 Linear model

Consider the model

$$y_i = \mu_i(\beta, \gamma(z_i)) + \epsilon_i = x_i'\beta + \gamma(z_i) + \epsilon_i \tag{1.8}$$

where ϵ_i's are independent and identically distributed with mean 0 and variance σ_ϵ^2. Here, $E(Y_i) = x_i'\beta + \gamma(z_i)$ with $\theta_i = h(x_i'\beta + \gamma(z_i)) = x_i'\beta + \gamma(z_i), h(.)$ being the identity function. Also, $var(Y_i) = \sigma_{ii} = \sigma_\epsilon^2,\ i = 1, \ldots, K$.

1.2.1.1 Estimation of non-parametric function $\gamma(z_0)$

For model (1.8), the quasi-likelihood function $Q(\mu_i, y_i)$ can be written as

$$Q(\mu_i, y_i) = \frac{y_i - \mu_i(\beta, \gamma(z_i))}{\sigma_\epsilon^2}$$

Then, the semi-parametric QL estimating equation for $\gamma(z_0)$ is

$$\sum_{i=1}^{K} w_i(z_0) \frac{\partial \mu_i(\beta, \gamma(z_0))}{\partial \gamma(z_0)} [\frac{y_i - \mu_i(\beta, \gamma(z_0))}{\sigma_\epsilon^2}] = 0 \tag{1.9}$$

where $w_i(z_0) = \frac{p_i(\frac{z_0 - z_i}{b})}{\sum_{i=1}^{K} p_i(\frac{z_0 - z_i}{b})}$, $p_i(.)$ being a kernel density function. For example, one may choose $p_i(\frac{z_0 - z_i}{b}) = \frac{1}{\sqrt{2\pi}b} \exp(\frac{-1}{2}(\frac{z_0 - z_i}{b})^2)$ with a suitable bandwidth b. Note that when $w_i(z_0) = 1$, this SQL equation further reduces to the well-known quasi-likelihood estimating equation [Wedderburn (1974)].

Since $\frac{\partial \mu_i(\beta, z_0)}{\partial \gamma(z_0)} = \frac{\partial [x_i'\beta + \gamma(z_0)]}{\partial \gamma(z_0)} = 1$, the SQL estimating equation (1.9) has the formula

$$\sum_{i=1}^{K} w_i(z_0) [\frac{y_i - x_i'\beta - \gamma(z_0)}{\sigma_\epsilon^2}] = 0 \tag{1.10}$$

$$\Rightarrow \sum_{i=1}^{K} w_i(z_0)(y_i - x_i'\beta) - \sum_{i=1}^{K} w_i(z_0)\gamma(z_0) = 0$$

yielding an estimate for the non-parametric function $\gamma(z)$ evaluated at $z = z_0$ as

$$\hat{\gamma}(z_0) = \frac{\sum_{i=1}^{K} w_i(z_0)(y_i - x_i'\beta)}{\sum_{i=1}^{K} w_i(z_0)} = \sum_{i=1}^{K} w_i(z_0)(y_i - x_i'\beta) \tag{1.11}$$

where $\sum_{i=1}^{K} w_i(z_0) = 1$. Now replacing z_0 in (1.11) with z_i, we write

$$\hat{\gamma}(z_i) = \sum_{j=1}^{K} w_j(z_i)(y_j - x_j'\beta) = \hat{y}_i - \hat{x}_i'\beta \tag{1.12}$$

where

$$\hat{y}_i = \sum_{j=1}^{K} w_j(z_i)y_j \quad \text{and} \quad \hat{x}_i = \sum_{j=1}^{K} w_j(z_i)x_j \tag{1.13}$$

Note that the estimator $\hat{\gamma}(z_i)$ in (1.12) is constructed for a given value of the regression parameter vector β. But, because in practice β is unknown and in fact it is the main parameter of interest, we provide the estimating equation for β in the following section. However, these formulas for $\hat{\gamma}(z_i)$ and $\hat{\beta}$ are already discussed in literature and for example, we refer to Severini and Staniswalis (1994), Speckman (1988) and Hastie and Tibshirani (1990).

1.2.1.2 Estimation of regression effects β

For linear models the QL estimator of β has a closed form expression. To derive the estimator, we first write $\mu_i(\beta, \hat{\gamma}(z_i)) = x_i'\beta + \hat{\gamma}(z_i)$ and compute

$$\begin{aligned} \frac{\partial \mu_i(\beta, \hat{\gamma}(z_i))}{\partial \beta} &= \frac{\partial}{\partial \beta}[x_i'\beta + \hat{\gamma}(z_i)] \\ &= \frac{\partial}{\partial \beta}[x_i'\beta + \hat{y}_i - \hat{x}_i'\beta] \\ &= (x_i - \hat{x}_i)', \end{aligned} \tag{1.14}$$

where $\hat{x}_i$ is given in (1.13). Similar to (1.5) we now write the QL estimating equation for β as

$$\sum_{i=1}^{K} (x_i - \hat{x}_i)'[\frac{y_i - x_i'\beta - \hat{\gamma}(z_i)}{\sigma_\epsilon^2}] = 0,$$

and by substituting $\hat{\gamma}(z_i) = \hat{y}_i - \hat{x}_i'\beta$ we obtain

$$\sum_{i=1}^{K}(x_i - \hat{x}_i)'[y_i - x_i'\beta - \hat{y}_i + \hat{x}_i'\beta] = \sum_{i=1}^{K}(x_i - \hat{x}_i)'[(y_i - \hat{y}_i) - (x_i - \hat{x}_i)'\beta] = 0,$$

yielding

$$\sum_{i=1}^{K}(x_i - \hat{x}_i)'(y_i - \hat{y}_i) = \sum_{i=1}^{K}(x_i - \hat{x}_i)'(x_i - \hat{x}_i)\beta.$$

It then follows that $\hat{\beta}$ has the closed form expression given by

$$\hat{\beta} = [\sum_{i=1}^{K}(x_i - \hat{x}_i)'(x_i - \hat{x}_i)]^{-1} \sum_{i=1}^{K}(x_i - \hat{x}_i)'(y_i - \hat{y}_i), \tag{1.15}$$

where $\hat{y}_i$ and $\hat{x}_i$ are given in equation (1.13). The above equation (1.15) is the same as in Severini and Staniswalis (1994) [eqn.(10), page. 503] with $D = I$, the identity matrix.

1.2.2 Count data model

There are many situations in practice where one becomes interested in analyzing count and binary data to understand the effect of covariates on the responses. Similar to normally distributed responses considered in the previous section, these responses also follow the exponential family. However, in the present semi-parametric setup we are interested in examining the regression effect when the mean response is assumed to consist of the fixed regression function as well as a non-parametric smooth function. For count responses, the Poisson density function $f(y_i)$ can be expressed as a special form of exponential family density (1.1) given by

$$f(y_i) = \frac{\exp(-\mu_i)\mu_i^{y_i}}{y_i!} = \frac{1}{y_i!} \exp[y_i \, log\mu_i - \mu_i]. \tag{1.16}$$

where $\theta_i = log\ \mu_i$ and $a(\theta_i) = \mu_i$.

Thus we write the Poisson mean and variance as

$$E(Y_i|x_i, z_i) = V(Y_i|x_i, z_i) = \mu_i(\beta, \gamma(z_i)).$$

where

$$\mu_i(\beta, \gamma(z_i)) = \exp(x_i'\beta + \gamma(z_i))$$

which is different than (1.8) under the linear case.

1.2.2.1 Estimation of non-parametric function $\gamma(z_0)$

The SQL estimating equation for $\gamma(z_0)$ in the count data has the form

$$\sum_{i=1}^{K} w_i(z_0) \frac{\partial \mu_i(\beta, \gamma(z_0))}{\partial \gamma(z_0)} [\frac{y_i - \mu_i(\beta, \gamma(z_0))}{\mu_i(\beta, \gamma(z_0))}] = 0 \tag{1.17}$$

where $\mu_i(\beta, \gamma(z_0)) = \exp(x_i'\beta + \gamma(z_0))$.

Because $\frac{\partial \mu_i(\beta, \gamma(z_0))}{\partial \gamma(z_0)} = \frac{\partial \exp(x_i'\beta + \gamma(z_0))}{\partial \gamma(z_0)} = \exp(x_i'\beta + \gamma(z_0))$, (1.17) reduces to

$$\sum_{i=1}^{K} w_i(z_0)[y_i - \exp(x_i'\beta + \gamma(z_0))] = 0 \tag{1.18}$$

and hence

$$\exp(\hat{\gamma}(z_0)) = \frac{\sum_{i=1}^{K} w_i(z_0) y_i}{\sum_{i=1}^{K} w_i(z_0) \exp(x_i'\beta)}.$$

The estimator for $\gamma(z)$ computed at $z = z_0$ under the Poisson model is then given by

$$\hat{\gamma}(z_0) = log \left(\frac{\sum_{i=1}^{K} w_i(z_0) y_i}{\sum_{i=1}^{K} w_i(z_0) \exp(x_i'\beta)} \right).$$

Thus for $z = z_i$ the estimator of and $\gamma(z)$ has the form

$$\hat{\gamma}(z_i) = log \left(\frac{\sum_{j=1}^{K} w_j(z_i) y_j}{\sum_{j=1}^{K} w_j(z_i) \exp(x_j'\beta)} \right). \tag{1.19}$$

1.2.2.2 Estimation of regression effects β

Unlike the linear models, the estimator of β has no explicit form under the Poisson count data model, and one has to estimate β by solving a non-linear equation iteratively. For this purpose, similar to (1.5), the QL estimating equation for β is

$$\sum_{i=1}^{K} \frac{\partial \mu_i(\beta, \hat{\gamma}(z_i))}{\partial \beta} [\frac{y_i - \mu_i(\beta, \hat{\gamma}(z_i))}{\mu_i(\beta, \hat{\gamma}(z_i))}] = 0 \tag{1.20}$$

where

$$\begin{aligned} \frac{\partial \mu_i(\beta, \hat{\gamma}(z_i))}{\partial \beta} &= \frac{\partial}{\partial \beta}[\exp(x_i'\beta + \hat{\gamma}(z_i))] \\ &= [\exp(x_i'\beta + \hat{\gamma}(z_i))]\ [x_i' + \frac{\partial \hat{\gamma}(z_i)}{\partial \beta}] \end{aligned} \tag{1.21}$$

with $\hat{\gamma}(z_i)$ as in (1.19). The derivative $\frac{\partial \hat{\gamma}(z_i)}{\partial \beta}$ is computed as

$$\begin{aligned} \frac{\partial \hat{\gamma}(z_i)}{\partial \beta} &= -\left[\frac{\sum_{j=1}^{K} w_j(z_i) y_j}{\sum_{j=1}^{K} w_j(z_i) \exp(x_j'\beta)}\right]^{-1} \frac{\sum_{j=1}^{K} w_j(z_i) y_j \ \sum_{j=1}^{K} w_j(z_i) \exp(x_j'\beta) x_j'}{[\sum_{j=1}^{K} w_j(z_i) \exp(x_j'\beta)]^2} \\ &= -\frac{\sum_{j=1}^{K} w_j(z_i) \exp(x_j'\beta) x_j'}{\sum_{j=1}^{K} w_j(z_i) \exp(x_j'\beta)}. \end{aligned} \tag{1.22}$$

Now by using (1.22) in (1.21) we write

$$\begin{aligned} \frac{\partial \mu_i(\beta, \hat{\gamma}(z_i))}{\partial \beta} &= [\exp(x_i'\beta + \hat{\gamma}(z_i))]\ [x_i' - \frac{\sum_{j=1}^{K} w_j(z_i) \exp(x_j'\beta) x_j'}{\sum_{j=1}^{K} w_j(z_i) \exp(x_j'\beta)}] \\ &= \mu_i(\beta, \hat{\gamma}(z_i))[x_i' - \frac{\sum_{j=1}^{K} w_j(z_i) \exp(x_j'\beta) x_j'}{\sum_{j=1}^{K} w_j(z_i) \exp(x_j'\beta)}]. \end{aligned}$$

Consequently, the estimating equation (1.20) leads to

$$\sum_{i=1}^{K} [x_i' - \frac{\sum_{j=1}^{K} w_j(z_i) \exp(x_j'\beta) x_j'}{\sum_{j=1}^{K} w_j(z_i) \exp(x_j'\beta)}][y_i - \tilde{\mu}_i] = 0$$

where $\tilde{\mu}_i = \exp(x_i'\beta + \hat{\gamma}(z_i))$. Now by defining

$$\hat{x}_i = \frac{\sum_{j=1}^{K} w_j(z_i) \exp(x_j'\beta) x_j}{\sum_{j=1}^{K} w_j(z_i) \exp(x_j'\beta)}, \tag{1.23}$$

we rewrite the estimating equation as

$$\sum_{i=1}^{K} (x_i - \hat{x}_i)' \ (y_i - \tilde{\mu}_i) = 0. \tag{1.24}$$

The estimating equation (1.24) can be solved iteratively using the well-known Newton-Raphson method. The iterative equation has the form

$$\begin{aligned} \hat{\beta}_{(r+1)} &= \hat{\beta}_{(r)} - [\frac{\partial}{\partial \beta'} \sum_{i=1}^{K} (x_i - \hat{x}_i)' \ (y_i - \tilde{\mu}_i)]^{-1} \ [\sum_{i=1}^{K} (x_i - \hat{x}_i) \ (y_i - \tilde{\mu}_i)] \\ &= \hat{\beta}_{(r)} + [\sum_{i=1}^{K} (x_i - \hat{x}_i)' \ \tilde{\mu}_i \ (x_i - \hat{x}_i)]^{-1} \ [\sum_{i=1}^{K} (x_i - \hat{x}_i) \ (y_i - \tilde{\mu}_i)] \end{aligned} \tag{1.25}$$

and is used to compute the final estimate $\hat{\beta}$ until convergence.

Severini and Staniswalis (1994, Example 2, page. 503) provided an estimate for $\gamma(z_i)$ under gamma distribution, which is similar, but different than (1.19). Hence for the estimation of β, we have provided the exact iterative equation in (1.25) under the Poisson case.

1.2.3 Binary data model

In the semi-parametric GLM set up for binary responses, the binary distribution is

$$f(y_i) = \mu_i^{y_i} \ (1 - \mu_i)^{1 - y_i}$$

which is a special case of the exponential family density (1.1) with

$$\theta_i = log\left(\frac{\mu_i}{1 - \mu_i}\right) \text{ and } a(\theta_i) = -log(1 - \mu_i).$$

In the partially specified regression case we consider $\theta_i = x_i'\beta + \gamma(z_i)$ and it then follows that

$$\mu_i = \frac{\exp(\theta_i)}{1+\exp(\theta_i)} \text{ and } a(\theta_i) = log(1+\exp(\theta_i))$$

yielding

$$E(Y_i|x_i, z_i) = a'(\theta_i) = \mu_i(\beta, \gamma(z_i))$$

and

$$V(Y_i|x_i, z_i) = a''(\theta_i) = \mu_i(\beta, \gamma(z_i))(1 - \mu_i(\beta, \gamma(z_i))).$$

1.2.3.1 Estimation of non-parametric function $\gamma(z_0)$

In the binary case, the SQL estimating equation for $\gamma(z)$ at $z = z_0$ is given by

$$\sum_{i=1}^{K} w_i(z_0) \frac{\partial \mu_i(\beta, \gamma(z_0))}{\partial \gamma(z_0)} \left[\frac{y_i - \mu_i(\beta, \gamma(z_0))}{\mu_i(\beta, \gamma(z_0))(1 - \mu_i(\beta, \gamma(z_0)))} \right] = 0, \tag{1.26}$$

where $\mu_i(\beta, \gamma(z_0)) = \frac{\exp(x_i'\beta + \gamma(z_0))}{1+\exp(x_i'\beta+\gamma(z_0))}$. Because,

$$\begin{aligned} \frac{\partial \mu_i(\beta, \gamma(z_0))}{\partial \gamma(z_0)} &= \frac{\exp(x_i'\beta + \gamma(z_0))}{1+\exp(x_i'\beta + \gamma(z_0))} \frac{1}{1+\exp(x_i'\beta + \gamma(z_0))} \\ &= \mu_i(\beta, \gamma(z_0))(1 - \mu_i(\beta, \gamma(z_0))), \end{aligned}$$

the estimating equation (1.26) reduces to

$$\sum_{i=1}^{K} w_i(z_0) \left[y_i - \mu_i(\beta, \gamma(z_0)) \right] = 0, \tag{1.27}$$

which is similar to (1.18). The difference lies in the formula for $\mu_i(\beta, \gamma(z_0))$.

1.2.3.2 Estimation of regression effects β

For the estimation of β, the QL estimating equation has the formula

$$\sum_{i=1}^{K} \frac{\partial \mu_i(\beta, \hat{\gamma}(z_i))}{\partial \beta} \left[\frac{y_i - \mu_i(\beta, \hat{\gamma}(z_i))}{\mu_i(\beta, \hat{\gamma}(z_i))(1 - \mu_i(\beta, \hat{\gamma}(z_i)))} \right] = 0 \tag{1.28}$$

where

$$\begin{aligned}\frac{\partial \mu_i(\beta, \hat{\gamma}(z_i))}{\partial \beta} &= \frac{\partial}{\partial \beta}\left[\frac{\exp(x_i'\beta + \hat{\gamma}(z_i))}{1+\exp(x_i'\beta + \hat{\gamma}(z_i))}\right] \\ &= \left[\frac{\exp(x_i'\beta + \hat{\gamma}(z_i))}{[1+\exp(x_i'\beta + \hat{\gamma}(z_i))]^2}\right]\ [x_i' + \frac{\partial \hat{\gamma}(z_i)}{\partial \beta}] \\ &= \mu_i(\beta, \hat{\gamma}(z_i))\ (1-\mu_i(\beta, \hat{\gamma}(z_i)))\ [x_i' + \frac{\partial \hat{\gamma}(z_i)}{\partial \beta}]\end{aligned}$$

The estimating equation in (1.28) then reduces to

$$\sum_{i=1}^{K}[x_i' + \frac{\partial \hat{\gamma}(z_i)}{\partial \beta}][y_i - \mu_i(\beta, \hat{\gamma}(z_i))] = 0 \tag{1.29}$$

Note that the estimating equation for $\gamma(.)$ in (1.27), and the estimating equation for β in (1.29) are the same as those in equations (6) and (8) respectively in Severini and Staniswalis (1994), and that these equations must be solved iteratively. However, there is a closed form expression for $\gamma(.)$ (1.19) in the Poisson case, whereas the estimating equation (1.27) for the binary case has to be solved iteratively. One needs to solve the estimating equation for β iteratively both in binary and in Poisson cases.

1.3 Generalized linear longitudinal models (GLLMs)

We have discussed the GLMs in independent set up in section 1.1 and its generalization to the independent semi-parametric set up in details in section 1.2. The purpose of this research is to study the model and inferences in the semi-parametric longitudinal data. For convenience, in this section, we now review the existing models and associated inferences in longitudinal set up.

In notation, let $y_i = (y_{i1}, \ldots, y_{it}, \ldots, y_{iT})'$ represent the response vector, where y_{it} is the response recorded at time t for the i th individual. Suppose that $x_{it} =$

$(x_{it1}, \ldots, x_{itv}, \ldots, x_{itp})'$ be the $p-$ dimensional covariate vector corresponding to the scalar y_{it}, and β be the $p-$ dimensional regression effects of x_{it} on y_{it} for all $i = 1, \ldots, K$, and all $t = 1, \ldots, T$. Since the same outcome is measured consecutively over time for each individual, the repeated responses of an individual are likely to be correlated. In this set up we assume that the response y_i marginally follows (1.1) but their joint distribution is difficult to write, especially for discrete responses. The mean and variance of the response are denoted by $\mu_{it}(\beta) = a'(\theta_{it}) = E[Y_{it}]$ and $var[Y_{it}] = a''(\theta_{it}) = \sigma_{itt}(\beta)$. Similar to (1.5), the QL estimating equation for the unknown regression parameter β can be written as

$$\begin{aligned} &\sum_{i=1}^{K}\sum_{t=1}^{T} \frac{\partial a'(\theta_{it})}{\partial \beta}[a''(\theta_{it})]^{-1}(y_{it} - a'(\theta_{it})) \\ &= \sum_{i=1}^{K}\sum_{t=1}^{T} \frac{\partial \mu_{it}(\beta)}{\partial \beta}[\sigma_{itt}(\beta)]^{-1}(y_{it} - \mu_{it}(\beta)) = 0 \end{aligned} \tag{1.30}$$

The QL estimating equation (1.30) is the same as the independence assumption based QL estimating equation and the solution of this estimating equation provides a consistent, but inefficient, estimate for β. This is because the observations from the same individual are correlated and (1.30) is written ignoring such correlations. As a remedy, one must take the correlations of longitudinal responses into account to achieve the desired efficiency of the regression estimates.

The relevant works in the field of longitudinal data analysis originated from Liang and Zeger (1986). The authors introduce an extension of GLM for independent data to the longitudinal setup and propose the generalized estimating equations (GEEs) to acquire consistent and efficient regression estimates involved in the GLLM model. The backbone of their methodology is based on a 'working' correlation matrix. Liang

and Zeger defined the GEE estimating equation as

$$\sum_{i=1}^{K} \frac{\partial \mu_i'(\beta)}{\partial \beta} V_i(\alpha)^{-1}(y_i - \mu_i(\beta)) = 0, \tag{1.31}$$

where $\mu_i(\beta) = (\mu_{i1}(\beta).\ldots, \mu_{it}(\beta), \ldots. \mu_{iT}(\beta))'$ is the mean vector of y_i and $V_i(\alpha) = A_i^{1/2} R_i(\alpha) A_i^{1/2}$ is the covariance matrix with $A_i = \text{diag}[\sigma_{i11}(\beta), \ldots. \sigma_{ijj}(\beta), \ldots \sigma_{iTT}(\beta)]$, $R_i(\alpha)$ is a 'working' correlation matrix, and α is the 'working' correlation parameter. Subsequent research in the longitudinal data analysis literature shows that, in several situations, these 'working' correlation based regression parameter estimates are inconsistent [Crowder (1995)]. Crowder showed that this consistency breakdown occurs due to the problem in estimating the so-called 'working' correlation parameter α. In cases where 'working' correlations are estimable, Sutradhar and Das (1999) showed that even if the estimator of α converges to a value, the GEE approach gives consistent estimators of the regression parameters, but these estimators may be less efficient than the regression estimators obtained based on the independence estimating equations approach. Sutradhar (2003) proposed a generalization of the QL estimation approach, where β is obtained by solving the generalized quasi-likelihood (GQL) estimating equation given by

$$\sum_{i=1}^{K} \frac{\partial \mu_i'(\beta)}{\partial \beta} \Sigma_i^{-1}(\rho)(y_i - \mu_i(\beta)) = 0, \tag{1.32}$$

where $\mu_i(\beta) = (\mu_{i1}(\beta), \ldots, \mu_{it}(\beta), \ldots, \mu_{iT}(\beta))'$ is the mean vector of y_i and $\Sigma_i(\rho) = A_i^{1/2} C_i(\rho) A_i^{1/2}$ is the covariance matrix with $A_i = \text{diag}[\sigma_{i11}(\beta), \ldots, \sigma_{ijj}(\beta), \ldots \sigma_{iTT}(\beta)]$, $C_i(\rho)$ is a general class of auto-correlations, and ρ is a correlation index parameter. The estimator $\hat{\beta}_{GQL}$ obtained by solving (1.32) is consistent and very efficient for β.

1.4 Semi-parametric GLLMs

In the above mentioned longitudinal studies, regression functions involved in the longitudinal model are fully specified. For example, in linear longitudinal set up $\mu_{it}(\beta)$ is expressed as $\mu_{it}(\beta) = x_{it}\beta$. This leads to parametric modelling of marginal longitudinal models [Gilmour, Anderson, and Rae (1985), Liang and Zegger (1986), Zeger and Liang (1986), Fitzmaurice, Laird and Rotnitzky (1993)]. However, there are situations where the regression functions involved in the model are partially specified, which leads to semi-parametric models in the longitudinal setup. In the linear longitudinal setup, the semi-parametric models have been studied by Severini and Wong (1992), Zeger and Diggle (1994), Moyeed and Diggle (1994), You and Chen (2007), Fan, Haung and Li (2007), Fan and Wu (2008), and Li (2011). Some of these studies used the 'working' correlations based GEE approach for the estimation of regression parameters, and the non-parametric function was estimated separately by using independence assumption [see Zeger and Diggle (1994)]. Other works such as Fan, Haung and Li (2007) assumed normality for the responses and used likelihood approach for the estimation. But the covariance matrix for the multivariate distribution was constructed based on the 'working' correlation matrix. There also exist some generalizations where heteroscedasticity is assumed among the responses at a given time.

The semi-parametric analysis has also been studied for (marginal) exponential family data by using the 'working' correlations based GEE approach. To be specific, we refer to Severini and Staniswalis (1994), Lin and Carroll (2001, 2001a) for this GEE based analysis. These studies estimate regression parameters and non-parametric

functions separately and GEE approaches has been used in both cases.

1.5 Objective of the thesis

The main objective of this thesis is to study the semi-parametric regression models when the repeated responses follow a non-stationary correlation model that belongs to a class of Gaussian-type ARMA correlation structures. The plan of the thesis is as follows.

In **Chapter 2**, we focus on the semi-parametric linear longitudinal model where a stationary correlation structure is used for inference. In the linear model setup, this type of stationary correlation structure is quite appropriate because the correlations under linear models do not depend on any covariates irrespective of whether the covariates are time dependent. Even though the semi-parametric analysis in the linear model setup for longitudinal data is a direct extension of the independence based semi-parametric analysis discussed in Section 1.2, a close look at the estimation problem (to be discussed in Chapter 2) reveals that the existing studies in the semi-parametric longitudinal setup did not incorporate the estimation effects of non-parametric function $\gamma(.)$ while estimating the main regression parameter β. Also, the existing studies have extended the 'working' correlations based GEE approach explained in (1.31) to the semi-parametric setup, which may not provide efficient regression estimates. To overcome these two problems, we revisit the inferences for the semi-parametric linear longitudinal models and provide appropriate estimating equations for efficient inferences by using (1) ARMA type class of auto-correlation structures, and (2) taking the the estimation effect of non-parametric function in

estimating β. We carry out a simulation study to examine the finite sample based efficiencies of the proposed semi-parametric GQL (SGQL) as well as various semi-parametric GEE (SGEE) approaches. The asymptotic distribution of the proposed estimator is also discussed.

In **Chapter 3**, we extend the semi-parametric linear longitudinal model discussed in chapter 2, to the discrete data setup. In particular, we consider semi-parametric models for longitudinal count and binary data. Note that some of the existing studies such as Lin and Carroll (2001) and Severini and Staniswalis (1994) deal with such models, but they mainly use the 'working' correlations based GEE approach. These studies do not appear to accommodate the estimation effect of the non-parametric function $\gamma(.)$ while estimating β. As far as the correlation structure is concerned, in our approach, we use the non-stationary correlation structures suggested by Sutradhar (2010) for both count and binary data. However, we do not discuss any diagnostic procedure for the identification of the non-stationary correlation structure but this can be done following the technique given in Sutradhar (2010, Section 4). Rather, we assume that the correlation structure involving the time dependent covariates are known and develop a semi-parametric GQL (SGQL) approach for the main regression parameters by taking the estimation effect of the non-parametric function as well as the longitudinal correlations into account. Analytical details for the SGQL approach for both count and binary data are also provided. For the comparison with the existing studies, the proposed SGQL estimating equation is written in two ways. First, a partially standardized SGQL (PSSGQL) approach is described where the covariance matrix involved in the estimating equation for β is free from the estimation effect of $\gamma(.)$. Second, a fully standardized SGQL (FSSGQL) approach is discussed in which

the estimation effect of $\gamma(.)$ is accommodated in the covariance matrix.

To examine the finite sample performance of the proposed SGQL approaches, we carry out several simulation studies in **Chapter 4** for the longitudinal count data. First we study the effect of ignoring the non-parametric function in estimating β using a naive GQL (NGQL) approach. Because the performance of the leading GEE based approaches did not adequately study the count data in the semi-parametric setup, we have made a detailed comparison of the proposed PSSGQL approach with the existing partially standardized semi-parametric GEE (PSSGEE) approaches in order to achieve effiecient inference methods. We also provide the simulation results for the proposed FSSGQL approach.

The thesis concludes in **Chapter 5**.

Chapter 2

Semi-parametric Linear Longitudinal Models

In this chapter, we revisit the semi-parametric analysis for linear longitudinal data collected over equi-spaced and unbalanced time points. However, we use general notations such that the regression function can be written for the responses collected over unequi-spaced time points, which accommodate the equi-spaced time data as an important special case. As far as the correlation structure for the repeated responses is concerned, we concentrate on equi-spaced time data only. Thus, as opposed to the notation y_{it} used in Section 1.3 to represent the response at time t $(t = 1, \ldots, T)$ from the i^{th} $(i = 1, \ldots, K)$ individual, we now use a general notation, namely, $y_{ij}(t_{ij})$ to denote the j^{th} $(j = 1, \ldots, n_i)$ response of the i^{th} individual at time t_{ij}. Here n_i denotes the total number of responses for the i^{th} individual collected over n_i time points. Further, for equi-spaced time data, the time points would satisfy the relationship $t_{ij} - t_{i,j-1} \equiv t_{i,j+1} - t_{i,j}$, for example.

Suppose that $y_i = (y_{i1}(t_{i1}), \ldots, y_{ij}(t_{ij}), \ldots, y_{in_i}(t_{in_i}))'$ denotes the $n_i \times 1$ vector of repeated responses for the i^{th} $(i = 1, \ldots, K)$ individual. Also suppose that these repeated responses are influenced by a smooth non-parametric function $\gamma(t_{ij})$, and a fixed and known $p \times n_i$ covariate matrix $X_i' = (x_{i1}(t_{i1}), \ldots, x_{ij}(t_{ij}), \ldots, x_{in_i}(t_{in_i}))$, $x_{ij}(t_{ij})$ being the $p-$dimensional covariate vector at time point t_{ij}. This type of repeated continuous data measured at time point t_{ij} is usually modelled as

$$\begin{aligned} y_{ij}(t_{ij}) &= x_{ij}'(t_{ij})\beta + \gamma(t_{ij}) + \epsilon_{ij}(t_{ij}) \\ &= \mu_{ij}(t_{ij}) + \epsilon_{ij}(t_{ij}), \end{aligned} \tag{2.1}$$

or equivalently

$$y_i = X_i\beta + \gamma(t_i) + \epsilon_i, \tag{2.2}$$

where $\gamma(t_i) = (\gamma(t_{i1}), \cdots, \gamma(t_{in_i}))'$ and $\epsilon_i = (\epsilon_{i1}(t_{i1}), \cdots, \epsilon_{ij}(t_{ij}), \cdots, \epsilon_{in_i}(t_{in_i}))'$. We assume, $E(\epsilon_i) = 0$ and $\text{var}(\epsilon_i) = \text{var}(Y_i) = \Sigma_i$.

Note that in (2.2), $\gamma(t_i)$ is not a subject specific non-parametric function as its construction requires only knowing $\gamma(t)$ at any time t [Zeger and Diggle (1994); Sneddon and Sutradhar (2004)]. To be specific, $\gamma(t_i)$ is used here to represent n_i components, each with the same non-parametric function but evaluated at n_i different time points for the i^{th} individual.

To develop an efficient estimation procedure it is important to consider the correlation structure of the repeated responses. Let $\rho_{|t_{ij}-t_{ik}|}$ denote the pairwise correlations between the two responses $y_{ij}(t_{ij}, t_{ik})$ for all $j \neq k; j, k = 1, \ldots, n_i$. The $n_i \times n_i$ correlation matrix for $y_i = (y_{i1}(t_{i1}), \ldots, y_{ij}(t_{ij}), \ldots, y_{in_i}(t_{in_i}))'$ is denoted by

$$C_i(\rho) = (\rho_{|t_{ij}-t_{ik}|}) : n_i \times n_i.$$

For the purpose of constructing a suitable estimating equation for β, it is necessary to obtain an estimate $\hat{C}_i(\rho)$ to compute $\hat{\Sigma}_i(\rho) = A_i^{\frac{1}{2}} \hat{C}_i(\rho) A_i^{\frac{1}{2}}$. However, in an experiment where an individual can report a response at any time that is, when $t_{ij} \neq t_{hj},\ i \neq h, i, h = 1, \ldots, K$, it is possible that in some situations the $C_i(\rho)$ matrices may have unbalanced dimensions. In other situations, it may happen that any two matrix $C_i(\rho)$ and $C_h(\rho)$ with $n_i = n_h$ may not be the same. In such cases, it is impossible to estimate $C_i(\rho)$ for i^{th} individual borrowing information from other (remaining) individuals. For this reason, many authors have written the estimating equations for β and $\gamma(.)$ for general case, that is, for unequi-spaced and unequal time for individuals, but the estimation for the correlation matrices was given for (1) $n_i = n$ for $i = 1, \ldots, K$, and (2) under the assumption that $C_i(\rho) = C(\rho)$, a constant and common matrix. For example, we refer to Lin and Carroll (2001, p. 1048) where $C_i(\rho)$ was estimated by

$$\hat{C}(\rho) = \frac{1}{K} \sum_{i=1}^{K} r_i r_i', \tag{2.3}$$

with $r_i = A_i^{-\frac{1}{2}}(y_i - X_i\hat{\beta} - \hat{\gamma}(t_i))$, where

$$A_i = \text{diag}[\sigma_{i11}(t_{i1}), \ldots, \sigma_{ijj}(t_{ij}), \ldots \sigma_{in_in_i}(t_{in_i})]$$

and $\sigma_{ijj}(t_{ij})$ is the variance of $\epsilon_{ij}(t_{ij})$.

Note that there are few difficulties with this correlation matrix (2.3) construction. This is because: (1) as the unbalanced $n_i \times n_i$ matrices $(r_i r_i')$ cannot be added from all individuals, $\hat{C}(\rho)$ computation is meaningful only when $n_i = n$, say. However, it is not understood how one may compute $C_i(\rho)$ needed for the construction of $\hat{\Sigma}_i$, when dimensions are not same (2) when a situation is considered where t_{ij}'s may be unequi-spaced, there is no reason to justify the use of $n_i = n$ for all i.

In the thesis, we concentrate on equi-spaced data and study the inferences for the regression effects in the semi-parametric setup by properly accommodating the longitudinal correlations for both continuous and discrete data. This type of data were used in Sutradhar (2010), but the author dealt only with a fixed (specified) regression function as opposed to a semi-parametric rergession function. As far as the correlation structure is concerned, following Sutradhar (2011), we assume that the repeated data follow a class of auto-correlation structures that accommodates Gaussian type all possible auto-regressive moving average of order r, s (ARMA(r, s)) correlation models with AR(1), MA(1), AR(2), MA(2), EQC (equi-correlations), as some special cases. Note that the AR(1), MA(1), and EQC structures for repeated data were also discussed in Liang and Zeger (1986), and subsequently these structures were used by Severini and Staniswallis (1994) in the semi-parametric longitudinal setup. Further note that in this approach it is not necessary that $n_i = n$ (balanced data) for all $i = 1, \ldots K$.

Specifically, we consider the correlation matrix $C_i(\rho)$ for the error vector ϵ_i in (2.2) as

$$C_i(\rho) = \begin{pmatrix} 1 & \rho_1 & \rho_2 & \ldots & \rho_{n_i-1} \\ \rho_1 & 1 & \rho_1 & \ldots & \rho_{n_i-2} \\ \vdots & & & \ldots & \vdots \\ \rho_{n_i-1} & \rho_{n_i-2} & & \ldots & 1 \end{pmatrix} \quad \text{for all i} = 1, 2, \ldots, K;$$

$$\Sigma_i(\rho) = \text{var}(Y_i) = A_i^{\frac{1}{2}} C_i(\rho) A_i^{\frac{1}{2}}, \tag{2.4}$$

where for $\ell = 1, \ldots, n_i - 1$, ρ_ℓ denotes the lag ℓ correlation between $\epsilon_{ij}(t_{ij})$ and $\epsilon_{i,j+\ell}(t_{i,j+\ell})$. We assume, however, that the variances are stationary and hence write

$A_i = \sigma^2 I_{n_i}$ where σ^2 is an unknown scalar constant, and I_{n_i} is the $n_i \times n_i$ identity matrix. The following examples demonstrate the correlation models that produce $C_i(\rho)$ in (2.4) in the linear model setup:

(i) **AR(1) model:**

$$\begin{aligned} \epsilon_{ij}(t_{ij}) &= \phi\ \epsilon_{i,j-1}(t_{i,j-1}) + a_{ij}(t_{ij}),\ |\phi| < 1, \\ & a_{ij}(t_{ij}) \overset{iid}{\sim} N(0, \sigma_a^2)\ \ \forall\ \mathrm{i} = 1, 2, \ldots, K; j = 1, \ldots, n_i, \end{aligned} \tag{2.5}$$

(ii) **MA(1) model:**

$$\begin{aligned} \epsilon_{ij}(t_{ij}) &= \theta\ a_{i,j-1}(t_{i,j-1}) + a_{ij}(t_{ij}),\ |\theta| < 1, \\ & a_{ij}(t_{ij}) \overset{iid}{\sim} N(0, \sigma_a^2)\ \ \forall\ \mathrm{i} = 1, 2, \ldots, K; j = 1, \ldots, n_i, \end{aligned} \tag{2.6}$$

and

(iii) **EQC model :**

$$\begin{aligned} \varepsilon_{ij}(t_{ij}) &= \varepsilon_{i0}(t_{i0}) + a_{ij}(t_{ij}), \\ & a_{ij}(t_{ij}) \overset{iid}{\sim} (0, \sigma_a^2),\ \ \varepsilon_{i0}(t_{i0}) \sim N(0, \tilde{\sigma}^2) \end{aligned} \tag{2.7}$$

The lag ℓ correlations (ρ_ℓ) between $\epsilon_{ij}(t_{ij})$ and $\epsilon_{i,j+\ell}(t_{i,j+\ell})$ for (2.4), (2.5) and (2.6) are

$$\rho_\ell = \phi^\ell, \ell = 1, \ldots, n_i - 1; \quad \rho_\ell = \begin{cases} \frac{\theta}{1+\theta^2}, & \text{for } \ell = 1 \\ 0, & \text{for } \ell = 2, 3, \ldots, n_i - 1, \end{cases} \quad \text{and}$$

$\rho_\ell = \zeta = \frac{\tilde{\sigma}^2}{\tilde{\sigma}^2 + \sigma_a^2}, \ell = 1, \ldots, n_i - 1$ respectively, and they satisfy the auto-correlation structure $C_i(\rho)$ in (2.4).

Note that even though the $C_i(\rho)$ matrix in (2.4) is written corresponding to n_i time points of the i^{th} individual, the exact structures for two individuals i and k,

with $n_i = n_k = n^*$, can be different when n_i time points do not overlap with n_k time points. In such a case, for $n = \max_i n_i,\ i = 1, \ldots, K$, a $n \times n$ correlation matrix is first computed and then $C_i(\rho)$ for the i^{th} individual is computed by deleting all rows and columns of the $n \times n$ matrix except those rows and columns corresponding to n_i time points. Similarly, $C_k(\rho)$ is constructed.

As far as the estimation of the regression effects is concerned, a 'working' correlations approach has been widely used both in fully specified and semi-parametric longitudinal setups, where one does not care about modelling the true correlation structure of the repeated responses. This approach is completely different than our parametric modelling of the true correlations, as it uses the general auto-correlation structure $C_i(\rho)$. Thus, $C_i(\rho)$ is not a working correlation matrix. Now, if $C_i(\rho)$ is treated as a working correlation matrix, and if the true structure belongs to the ARMA(p,q) class of auto-correlations, then logically such a 'working' selection would be efficient as it becomes a parametric model. The 'working' correlations approach, however, is used for any unknown true correlation structures with the hope that it does not loose much efficiency even if the 'working' structure is misspecified. But, it has been demonstrated by Sutradhar and Das (1999) [see also Sutradhar (2011)] in the complete longitudinal setup, for example, that even if the true correlations belong to an auto-correlations class, the use of a 'working' correlation structure such as the equi-correlations structure may produce inefficient regression estimates compared to the simpler 'independence' assumption based estimates. Moreover, in the 'working' correlation approach there is no guidance of preferring one correlation structure over the other, which frequently leads one to use either 'working' equi correlations or independence or unstructured correlations [Lin and Carroll (2001), Severini and

Staniswallis (1994)]. This type of individual specified 'working' correlation structures, however, may lead to inefficient regression estimates as compared to the $C_i(\rho)$ based parametric modelling when the true correlations belong to the aforementioned general auto-correlations class. For this reason, as opposed to the 'working' correlations based approaches, we use an auto-correlation structure (2.4) based semi-parametric generalized quasi-likelihood (SGQL) approach that always produces the same, or more efficient, regression estimates than the 'working' correlations based semi-parametric approaches.

We first review semi-parametric GEE (SGEE) approaches. It is well known that when y_i is influenced by fixed covariates X_i only, the generalized least square (GLS) estimator given by

$$\hat{\beta}_{GLS} = [\sum_{i=1}^{K} X_i' \hat{\Sigma}_i^{-1}(\rho) X_i]^{-1} \sum_{i=1}^{K} X_i' \hat{\Sigma}_i^{-1}(\rho) y_i \tag{2.8}$$

is the best linear unbiased estimate (BLUE) [Rao (1973, Section 4a.2), Amemiya (1985, Section 6.1.3)] for the regression parameter vector β within a class of linear unbiased estimators. However, when the response vector y_i is influenced by both fixed covariates X_i and an unspecified non-parametric vector function $\gamma(t_i) = (\gamma(t_{i1}), \cdots . \gamma(t_{in_i}))'$ as in (2.2), this GLS estimator (2.8) is biased and hence inconsistent for the true regression parameter β. Existing studies [see Severini and Staniswalis (1994), Lin and Carroll (2001)] estimate the non-parametric function consistently by using the kernel-based approaches, but the specified regression function is estimated by solving a working' correlations based SGEE approach. A close look at the derivation of the SGEE reveals that the gradient function used in constructing the estimating equation is correctly computed by taking the estimation effect of $\gamma(t)$ into

account, but the covariance matrix used in the estimating equation is constructed by ignoring the estimation effect of $\gamma(t)$ and this makes the SGEE partially standardized. As opposed to this partially standardized SGEE (PSSGEE) approach, we propose a fully standardized semi-parametric generalized quasi-likelihood (FSSGQL) approach where both the gradient function and the covariance matrix are constructed by taking the estimation effect for $\gamma(t)$ into account. Thus, FSSGQL approach provides more efficient regression estimates. The efficiency gain by the FSSGQL approach compared to the PSSGEE approaches is further demonstrated in Section 2.3 through an empirical study.

2.1 Existing semi-parametric estimation methods

2.1.1 PSSGEE approach

It follows from the model (2.1)-(2.2) that the mean response is given by

$$E[Y_{ij}(t_{ij})] = \mu_{ij}(t_{ij}) = x'_{ij}(t_{ij})\beta + \gamma(t_{ij}), \tag{2.9}$$

where β is the fixed regression effects, and $\gamma(t_{ij})$ is a non-parametric smooth function of time. Authors such as Zeger and Diggle (1994) consider

$$\text{cov}(Y_i) = \sigma^2 R_i(\alpha),$$

where $R_i(\alpha)$ is a 'working' correlation matrix used for the unknown true correlation matrix and α is the 'working' correlation parameter. The commonly used $R_i(\alpha)$ are: (a) the unstructured form $R_i(\alpha) = (r_{i,jk}(\alpha))$ with $r_{i,jk}(\alpha) = \alpha_{|t_{ij}-t_{ik}|}$ [Zeger and Diggle (1994), Lin and Carroll (2001)]; (b) equi correlations form $R_i(\alpha) = \alpha I_{n_i}$, and (c) independence form $R_i(\alpha) = I_{n_i}$ [Lin and Carroll (2001), Severini and Staniswalis (1994)].

Thus, for the semi-parametric linear longitudinal model, one needs to estimate the fixed regression effects β, the non-parametric smooth function $\gamma(t_{ij})$, the variance parameter σ^2, and the 'working' correlation matrix $R_i(\alpha)$. All these parameters and function have to be solved iteratively until convergence.

Even though β and $\gamma(t)$ together constitute the regression function, their joint estimation may be difficult. Thus, in the existing literature they are estimated marginally by using separate estimating equations [Zeger and Diggle (1994), Severini and Staniswallis (1994), and Lin and Carroll (2001)]. This makes it simpler, for example, to use the'working' independence approach for consistent estimation of $\gamma(t)$ [Zeger and Diggle (1994, Section 3.1)], and a suitable correlation structure based approach for efficient estimation of the main regression parameter β. Following this strategy, in the next section, we briefly explain how one can construct the 'working' independence assumption based estimating equation for $\gamma(t)$.

2.1.1.1 Estimation of non-parametric function

QL approach

Non-parametric kernel regression is widely used for the estimation of $\gamma(t)$. A 'working' independence assumption based unbiased estimating function is weighted by using suitable kernel weights, and the resulting semi-parametric estimating equation is then solved for $\gamma(t)$. The SQL estimating equation for $\gamma(t_0)$ is

$$\sum_{i=1}^{K}\sum_{j=1}^{n_i} w_{ij}(t_0)\frac{\partial \mu_{ij}}{\partial \gamma(t_0)}\frac{(y_{ij}-\mu_{ij})}{\sigma^2} = 0 \tag{2.10}$$

where $w_{ij}(t_0) = \frac{p_{ij}(\frac{t_0-t_{ij}}{b})}{\sum_{i=1}^{K}\sum_{j=1}^{n_i} p_{ij}(\frac{t_0-t_{ij}}{b})}$, $p_{ij}(.)$ is a suitable kernel function and b is the bandwidth parameter. When $w_{ij}(t_0) = 1$, the SQL equation (2.10) reduces to the

standard QL estimating equation [Wedderburn (1974), McCullagh (1983)]. Authors such as Sneddon and Sutradhar(2004), Zeger and Diggle (1994) and You and Chen (2007) have used such an estimating equation in the linear semi-parametric model setup. Because $\mu_{ij}(t_{ij}) = x'_{ij}(t_{ij})\beta + \gamma(t_{ij})$ by (2.9), the solution of the SQL estimating equation (2.10), in terms of known β, is

$$\hat{\gamma}(t_{ij}) = \hat{y}_{ij} - \hat{x}'_{ij}(t_{ij})\beta,$$

where

$$\hat{y}_{ij} = \sum_{h=1}^{K}\sum_{u=1}^{n_h} w_{hu}(t_{ij})y_{hu} \text{ and } \hat{x}'_{ij}(t_{ij}) = \sum_{h=1}^{K}\sum_{u=1}^{n_h} w_{hu}(t_{ij})x'_{hu}(t_{hu})$$

with $\sum_{h=1}^{K}\sum_{u=1}^{n_h} w_{hu}(t_{ij}) = 1$. This formula will be exploited in the next section for the estimation of β.

A GEE approach

Severini and Staniswalis (1994) [see also Wang, Carroll and Lin (2005)] solved certain 'working' correlations based semi-parametric GEE for the estimation of $\gamma(t)$.

Lin and Carroll (2001) considered a 'working' correlation based GEE estimating equation to estimate $\gamma(t)$. They considered an arbitrary linear function in time, that is, $\gamma(t_{ij}) = \alpha_0 + \alpha_1\frac{(t_{ij}-t)}{b}$, where $\tilde{\alpha} = (\alpha_0, \alpha_1)'$ is a 2×1 vector of unknown parameters and b denotes the bandwidth parameter. The regression function, $\mu_i(X_i, t) = [\mu_{i1}(t_{i1}), \ldots, \mu_{in_i}(t_{in_i})]'$ where $\mu_{ij}(t_{ij}) = x'_{ij}\beta + \gamma(t_{ij})$. Lin and Carroll used the following two kernel estimation equations (symmetric and asymmetric) for the estimation of $\gamma(t)$

$$\sum_{i=1}^{K} \frac{\partial \mu'_i(X_i,t)}{\partial \tilde{\alpha}} \, [var(Y_i)]^{-1} \, W_{ib}(t) \, (Y_i - \mu_i(X_i,t)) = 0$$

$$\sum_{i=1}^{K} T'_i(t) \, \Delta_i(X_i,t) \, [var(Y_i)]^{-1} \, W_{ib}(t) \, (Y_i - \mu_i(X_i,t)) = 0, \qquad (2.11)$$

where $T_i(t)$ is the $n_i \times 2$ design matrix with j^{th} row $\{1, \frac{(T_{ij}-t)}{b}\}$, $\Delta_i = I_{n_i}$, and $W_{ib}(t)$ is the kernel weight matrix. For simplicity we write $W_i(t)$ for $W_{ib}(t)$. The kernel weight matrix $W_i(t)$ is then defined as

$$W_i(t_{i1}, \ldots, t_{in_i}) = \begin{pmatrix} w_i'(t_{i1}) \\ \vdots \\ w_i'(t_{ij}) \\ \vdots \\ w_i'(t_{in_i}) \end{pmatrix} \tag{2.12}$$

where $w_i'(t_{ij}) = [w_{i1}(t_{ij}), \ldots, w_{iu}(t_{ij}), \ldots, w_{in_i}(t_{ij})]$ with $w_{iu}(t)$ at a given time t given in (2.10). Furthermore, using a 'working' correlation $R_i(\alpha)$ the authors considered $\text{var}(Y_i) = A_i^{\frac{1}{2}} R_i(\alpha) A_i^{\frac{1}{2}}$ with $A_i = \text{diag}[\sigma_{i11}(t_{i1}), \ldots, \sigma_{in_in_i}(t_{in_i})]$.

2.1.1.2 Estimation of regression effects

Using $\hat{\gamma}(t_{ij}) = \hat{y}_{ij} - \hat{x}_{ij}'(t_{ij})\beta$ in (2.1),

$$\begin{aligned} y_{ij}(t_{ij}) &= x_{ij}'(t_{ij})\beta + \hat{\gamma}(t_{ij}) + \epsilon_{ij}^*(t_{ij}) \\ &= x_{ij}'(t_{ij})\beta + \hat{y}_{ij}(t_{ij}) - \hat{x'}_{ij}(t_{ij})\beta + \epsilon_{ij}^*(t_{ij}) \end{aligned} \tag{2.13}$$

where $\epsilon_{ij}^*(t_{ij})$ is a new error component. This $\epsilon_{ij}^*(t_{ij})$ is different from $\epsilon_{ij}(t_{ij})$ because some errors are induced by replacing $\gamma(t_{ij})$ with its estimate $\hat{\gamma}(t_{ij})$ in the model (2.2). The marginal properties of the new error component are discussed in Section 2.2.

Now for all elements of the i^{th} individual we use (2.13) and following the notation in (2.2), we write

$$y_i - \hat{y}_i = (X_i - \hat{X}_i)\beta + \epsilon_i^* \tag{2.14}$$

where

$$\hat{y}_i = \sum_{h=1}^{K} W_h(t_{i1}, \ldots, t_{in_i}) y_h, \text{ and } \hat{X}_i = \sum_{h=1}^{K} W_h(t_{i1}, \ldots, t_{in_i}) X_h \tag{2.15}$$

with $W_h(t_{i1}, \ldots, t_{in_i})$, a $n_i \times n_h$ kernel weights matrix defined in a similar way as (2.12).

Severini and Staniswalis((1994), eqns. (17) and (18)) and You and Chen (2007, Section 4.1) [see also Lin and Carroll (2001)], use the PSSGEE estimation approach, where the estimating equation has the form

$$\sum_{i=1}^{K} \frac{\partial \mu_i'}{\partial \beta} \left[var(Y_i)\right]^{-1} (y_i - \mu_i) = 0,$$

which for the linear model (2.14) leads to

$$\begin{aligned} \hat{\beta}_{PSSGLS} &= [\sum_{i=1}^{K} (X_i - \hat{X}_i)' \left[var(Y_i)\right]^{-1} (X_i - \hat{X}_i)]^{-1} \\ &\times \sum_{i=1}^{K} (X_i - \hat{X}_i)' \left[var(Y_i)\right]^{-1} (y_i - \hat{y}_i), \end{aligned} \tag{2.16}$$

with $var(Y_i) = \Sigma_i = A_i^{\frac{1}{2}} R_i(\alpha) A_i^{\frac{1}{2}}$, $R_i(\alpha)$ being a 'working' correlation matrix. When (2.16) is examined in light of (2.14), PSSGEE estimator in (2.16) is constructed using an incorrect weight matrix var(Y_i), whereas the correct covariance matrix should have been $var(Y_i - \hat{Y}_i)$.

2.1.1.3 Estimation of the 'working' correlation parameter α

The 'working' correlation parameter α has a definition problem [Crowder (1995)]. Suppose that a 'working' correlation estimate $\hat{\alpha}$ under an assumed 'working' correlation model is computed. This estimate usually does not converge to α as the data used for its computation may follow a different model. Thus, $\hat{\alpha}$ converges to α_0,

say. which is different than α [Sutradhar and Das (1999)]. As far as the formula for $\hat{\alpha}$ is concerned, it is developed based on the method of moments following the assumed 'working' correlation structure. For example, if a user decides to use an equi-correlation matrix as the 'working' correlation structure for all K individuals, then the estimate would satisfy the estimating equation

$$\sum_{i=1}^{K}\sum_{j\neq u}^{n_i}(\tilde{y}_{ij}\tilde{y}_{iu}-\alpha)=0, \tag{2.17}$$

[Liang and Zeger (1986), Sutradhar (2011, Section 6.4.3)] where

$$\tilde{y}_{ij}=\frac{y_{ij}-x_{ij}'\hat{\beta}-\hat{\gamma}(t_{ij})}{\hat{\sigma^2}},$$

with

$$\hat{\sigma^2}=\sum_{i=1}^{K}\sum_{j=1}^{n_i}(y_{ij}-x_{ij}'\hat{\beta}-\hat{\gamma}(t_{ij}))^2/\sum_{i=1}^{K}n_i.$$

Similarly, for the estimation of a 'working' unstructured correlation matrix, one uses the moment estimating formula

$$\hat{R}_i(\alpha)=\frac{1}{K\hat{\sigma}^2}\sum_{i=1}^{K}r_i r_i^{'} \tag{2.18}$$

[Lin and Carroll (2001)] where $r_i=(r_{i1},r_{i2},\ldots,r_{in_i})'$ is the vector of residuals with $r_{ij}=y_{ij}-x_{ij}^{'}\hat{\beta}-\hat{\gamma}(t_{ij})$.

2.1.2 Partially standardized semi-parametric heteroscedastic GEE (PSSHGEE) approach

Fan and Wu (2008) [see also Fan, Huang and Li (2007)] examined the semi-parametric varying-coefficient partially linear regression models and proposed a difference-based method to estimate the mean function. The authors computed the covariance function

of the longitudinal model using a quasi-maximum likelihood approach for the purpose of prediction and found that prediction is not sensitive to the correlation structure. However, these covariance estimates were not used for the estimation of the main parameters β and $\gamma(t)$. Fan et al (2007, Section 2.1), and Fan and Wu (2008, eqn. (1)) estimated the non-parametric function by using a similar SQL estimate for $\gamma(t)$ and by using time dependent variances denoted by $\sigma^2(t)$ at a given time t. For the estimation of the regression effects β, they have used different 'working' correlation structures in the PSSGEE based estimate given by (2.16). Fan and Wu (2008, eqn. (6)) used the ordinary least squares (OLS) technique which is the same as using (2.16) with correlation matrix $R_i(\alpha) = I_{n_i}$, ignoring correlations. For a given t, the heteroscedasticity, i.e., the time dependent variances were computed by

$$\sigma^2(t) = \frac{\sum_{i=1}^{K} \sum_{j=1}^{n_i} r_{ij}^2(t) w_{ij}(t)}{\sum_{i=1}^{K} \sum_{j=1}^{n_i} w_{ij}(t)}, \tag{2.19}$$

where $r_{ij}(t) = y_{ij}(t) - x'_{ij}(t)\hat{\beta} - \hat{\gamma}(t)$, and $w_{ij}(t)$ are defined as in (2.10). Thus, for the estimation of β by (2.16), Fan and Wu (2008) use $\Sigma_i(\alpha) = A_i = \text{diag}[\sigma^2(t_{i1}), \ldots, \sigma^2(t_{in_i})]$. We refer this independence assumption-based PSSHGEE approach as PSSHGEE(I) and the corresponding estimator is denoted by $\hat{\beta}_{PSSHGEE(I)}$.

The estimation of $\gamma(t)$ and $\sigma^2(t)$ is similar in both Fan et al (2007) and Fan and Wu (2008). However, for β estimation by (2.16), Fan et al (2007) assumed that the error vector ϵ_i in (2.2) follow a multivariate normal distribution with a 'working' correlation matrix $R_i(\alpha)$, and estimated the 'working' correlation parameter α by maximizing the normal likelihood [Fan et al (2007, eqns. (2)-(3))]. This estimator may be referred to as the PSSHGEE estimator. We include this approach in our empirical efficiency comparison in Section 2.3, but compute the lag correlations by

the moment approach, which does not require any normality assumption.

2.2 Proposed FSSGQL approach

2.2.1 Estimation of non-parametric function

We consider the independence assumption based QL estimating equation to estimate the non-parametric function $\gamma(.)$. The SQL estimating equation for $\gamma(t_0)$ is

$$\sum_{i=1}^{K}\sum_{j=1}^{n_i} w_{ij}(t_0)\frac{\partial \mu_{ij}}{\partial \gamma(t_0)}\frac{(y_{ij}-\mu_{ij})}{\sigma^2} = 0 \tag{2.20}$$

Using the Nadaraya-Watson kernel regression method [Nadaraya (1964) and Watson (1964)], the weights are calculated as $w_{ij}(t_0) = \frac{p_{ij}(\frac{t_0-t_{ij}}{b})}{\sum_{i=1}^{K}\sum_{j=1}^{n_i} p_{ij}(\frac{t_0-t_{ij}}{b})}$ such that $\sum_{i=1}^{K}\sum_{j=1}^{n_i} w_{ij}(t_0) = 1$. The kernel function is chosen to be $p_i(\frac{t_0-t_{ij}}{b}) = \frac{1}{\sqrt{2\pi}b}\exp(\frac{-1}{2}(\frac{t_0-t_{ij}}{b})^2)$ and b is the bandwidth parameter. Since $\frac{\partial \mu_{ij}}{\partial \gamma(t_0)} = 1$, (2.20) reduces to

$$\sum_{i=1}^{K}\sum_{j=1}^{n_i} w_{ij}(t_0)(y_{ij}-\mu_{ij}) = 0$$

yielding $\hat{\gamma}(t_0) = \hat{y}_{ij} - \hat{x}'_{ij}\beta$, provided β is known or estimated.

Thus, $\hat{\gamma}(t_0)$ at $t_0 = t_{ij}$ is

$$\hat{\gamma}(t_{ij}) = \hat{y}_{ij} - \hat{x}'_{ij}(t_{ij})\beta, \tag{2.21}$$

where

$$\hat{y}_{ij} = \sum_{h=1}^{K}\sum_{u=1}^{n_h} w_{hu}(t_{ij})y_{hu} \text{ and } \hat{x}'_{ij}(t_{ij}) = \sum_{h=1}^{K}\sum_{u=1}^{n_h} w_{hu}(t_{ij})x'_{hu}(t_{hu})$$

with $\sum_{h=1}^{K}\sum_{u=1}^{n_h} w_{hu}(t_{ij}) = 1$.

2.2.2 Estimation of β

Recall from (2.14) that β has to be estimated from the model

$$y_i - \hat{y}_i = (X_i - \hat{X}_i)\beta + \epsilon_i^*.$$

Let $E(\epsilon^*) = \mu_i$, and $var(\epsilon^*) = var(Y_i - \hat{Y}_i) = \Sigma_i^*$. One may then estimate β using the GLS approach by minimizing the generalized squared distance function

$$\sum_{i=1}^{K} (\epsilon_i^* - \mu_i^*)' \; \Sigma_i^* \; (\epsilon_i^* - \mu_i^*)$$

$$= \sum_{i=1}^{K} (y_i - \hat{y}_i - (X_i - \hat{X}_i)\beta - \mu_i^*)' \; \Sigma_i^* \; (y_i - \hat{y}_i - (X_i - \hat{X}_i)\beta - \mu_i^*), \qquad (2.22)$$

with respect to β.

In the existing PSSGEE approach, the estimating equation for β was constructed by using $var(Y_i)$ or its working estimate $V_i(\alpha)$ instead of $var(Y_i - \hat{Y}_i)$.

Computation of μ_i^*

To compute $\mu_i^* = E(\epsilon^*) = E(Y_i - \hat{Y}_i - (X_i - \hat{X}_i)\beta)$, we first calculate

$$\begin{aligned} E[\hat{Y}_i] &= \sum_{h=1}^{K} W_h(t_{i1}, \ldots, t_{in_i}) E[Y_h] \\ &= \sum_{h=1}^{K} W_h(t_{i1}, \ldots, t_{in_i})[X_h\beta + \gamma(t_h)] \\ &= \hat{X}_i\beta + \sum_{h=1}^{K} W_h(t_{i1}, \ldots, t_{in_i})\gamma(t_h). \end{aligned} \qquad (2.23)$$

Hence

$$\begin{aligned} E[\epsilon_i^*] &= E(Y_i - \hat{Y}_i) - (X_i - \hat{X}_i)\beta \\ &= \gamma(t_i) - \sum_{h=1}^{K} W_h(t_{i1}, \ldots, t_{in_i})\gamma(t_h) = \mu_i^* \end{aligned} \qquad (2.24)$$

The kernel weights involved in (2.24) are chosen such that for fixed t_{ij},

$$w_{hl}(t_{ij}) \to 1 \text{ as } |t_{hl} - t_{ij}| \to 0$$
$$w_{hl}(t_{ij}) \to 0 \text{ as } |t_{hl} - t_{ij}| \text{ goes away from 0,} \tag{2.25}$$

satisfying $\sum_{h=1}^{K} \sum_{l=1}^{n_h} w_{hl}(t_{ij}) = 1$. Thus, for such selection of kernel weights $\mu_i^* \to 0$ and bias will be negligible.

Computation of Σ_i^*

We compute the $var(Y_i - \hat{Y}_i)$ as follows

$$\begin{aligned}
\Sigma_i^* &= var(Y_i - \hat{Y}_i) \\
&= var(Y_i) + var(\hat{Y}_i) - 2\ \mathrm{cov}(Y_i, \hat{Y}_i) \\
&= \Sigma_i(\rho) + var\{\sum_{h=1}^{K} W_h(t_{i1}, \ldots, t_{in_i}) Y_h\} - 2\ cov\{Y_i, \sum_{h=1}^{K} W_h(t_{i1}, \ldots, t_{in_i}) Y_h\} \\
&= \Sigma_i(\rho) + \sum_{h=1}^{K} \sum_{h'=1}^{K} W_h(t_{i1}, \ldots, t_{in_i}) cov(Y_h, Y_{h'}) W_{h'}(t_{i1}, \ldots, t_{in_i}) \\
&\quad -2\{W_i(t_{i1}, \ldots, t_{in_i}) var(Y_i) + \sum_{h \neq i}^{K} W_h(t_{i1}, \ldots, t_{in_i}) cov(Y_i, Y_h)\}
\end{aligned} \tag{2.26}$$

where $\Sigma_i(\rho) = A_i^{\frac{1}{2}} C_i(\rho) A_i^{\frac{1}{2}}$ with $C_i(\rho)$ as given in (2.4). Note that ρ in $C_i(\rho)$ is expressed as $\rho \equiv (\rho_1 \ldots, \rho_{n_i})$. Because y_i's are independent for $i = 1, \ldots K$, the formula in (2.26) for Σ_i^* reduces to

$$\Sigma_i^* = \Sigma_i(\rho) + \sum_{h=1}^{K} W_h(t_{i1}, \ldots, t_{in_i}) \Sigma_h(\rho) W_h'(t_{i1}, \ldots, t_{in_i}) - 2W_i(t_{i1}, \ldots, t_{in_i}) \Sigma_i(\rho) \tag{2.27}$$

Under the limiting conditions in (2.25), Σ_i^* reduces to $\Sigma_i(\rho)$. Nevertheless, in practice using the correct covariance Σ_i^* in place of $\Sigma_i(\rho)$ is bound to provide more efficient estimate for β.

Minimization of the distance function in (2.22) is equivalent to solve the GLS estimating equation

$$\sum_{i=1}^{K} \frac{\partial[(X_i - \hat{X}_i)\beta + \mu_i^*]'}{\partial\beta}[\Sigma_i^*]^{-1} \{(y_i - \hat{y}_i) - (X_i - \hat{X}_i)\beta - \mu_i^*\} = 0, \qquad (2.28)$$

for β. In the linear model, the GLS estimating equation is same as the GQL estimating equation and for this reason, and to be uniform with the notations in the next chapters, we refer to the estimating equation obtained from (2.28) as the fully standardized semi-parametric GQL (FSSGQL) estimating equation for β. The solution of (2.28) is given by

$$\begin{aligned} \hat{\beta}_{FSSGQL} &= \{\sum_{i=1}^{K}(X_i - \hat{X}_i)' (\Sigma_i^*)^{-1} (X_i - \hat{X}_i)\}^{-1} \\ &\times \sum_{i=1}^{K}(X_i - \hat{X}_i)' (\Sigma_i^*)^{-1} (y_i - \hat{y}_i - \mu_i^*). \end{aligned} \qquad (2.29)$$

Note that the difference between Σ_i in (2.16) and Σ_i^* in (2.29) may not be negligible in practice. It depends on the choice of the kernel weights.

2.2.2.1 Basic properties of $\hat{\beta}_{FSSGQL}$

Unbiasedness of $\hat{\beta}_{FSSGQL}$:

In the case where μ_i^* is known, it can be shown as follows that the FSSGQL

estimator $\hat{\beta}_{FSSGQL}$ is unbiased for β.

$$\begin{aligned} E(\hat{\beta}_{FSSGQL}) &= \{\sum_{i=1}^{K}(X_i - \hat{X}_i)' \ (\Sigma_i^*)^{-1} \ (X_i - \hat{X}_i)\}^{-1} \\ &\quad \times E\left[\sum_{i=1}^{K}(X_i - \hat{X}_i)' \ (\Sigma_i^*)^{-1} \ (y_i - \hat{y}_i - \mu_i^*)\right] \\ &= \{\sum_{i=1}^{K}(X_i - \hat{X}_i)' \ (\Sigma_i^*)^{-1} \ (X_i - \hat{X}_i)\}^{-1} \\ &\quad \times \sum_{i=1}^{K}(X_i - \hat{X}_i)' \ (\Sigma_i^*)^{-1} \ \{E(y_i - \hat{y}_i - \mu_i^*)\} \\ &= \{\sum_{i=1}^{K}(X_i - \hat{X}_i)' \ (\Sigma_i^*)^{-1} \ (X_i - \hat{X}_i)\}^{-1} \\ &\quad \times \sum_{i=1}^{K}(X_i - \hat{X}_i)' \ (\Sigma_i^*)^{-1} \ \{(X_i - \hat{X}_i)\beta + \mu_i^* - \mu_i^*\} \\ &= \beta, \end{aligned} \tag{2.30}$$

because $E(Y_i - \hat{Y}_i) = (X_i - \hat{X}_i)\beta + \mu_i^*$.

If the kernel weights are chosen satisfying the limiting conditions (2.25), μ_i^* in (2.24) tends to zero. Nevertheless, one may still like to estimate μ_i^* for the computation of $\hat{\beta}_{FSSGQL}$ by (2.29). This may be done by using the SQL estimate of $\gamma(.)$ in (2.24). Hence, one obtains

$$\hat{\mu}_i^* = \hat{\gamma}(t_i) - \sum_{h=1}^{K} W_h(t_{i1}, \ldots, t_{in_i})\hat{\gamma}(t_h)$$

where $\hat{\gamma}(.)$, the SQL estimator of $\gamma(.)$ computed by (2.20) is consistent for $\gamma(.)$.

We now compute the bias in estimating β by (2.29) when μ_i^* is replaced by $\hat{\mu}_i^*$.

For the purpose, we calculate $E(\hat{\beta}_{FSSGQL})$ as follows.

$$\begin{aligned}
E(\hat{\beta}_{FSSGQL}) &= \{\sum_{i=1}^{K}(X_i - \hat{X}_i)' \ (\Sigma_i^*)^{-1} \ (X_i - \hat{X}_i)\}^{-1} \\
&\quad \times \sum_{i=1}^{K}(X_i - \hat{X}_i)' \ (\Sigma_i^*)^{-1} \ \{E(y_i - \hat{y}_i - \hat{\mu}_i^*)\} \\
&= \{\sum_{i=1}^{K}(X_i - \hat{X}_i)' \ (\Sigma_i^*)^{-1} \ (X_i - \hat{X}_i)\}^{-1} \\
&\quad \times \sum_{i=1}^{K}(X_i - \hat{X}_i)' \ (\Sigma_i^*)^{-1} \ \{(X_i - \hat{X}_i)\beta + \mu_i^* - E(\hat{\mu}_i^*)\} \\
&= \beta + \{\sum_{i=1}^{K}(X_i - \hat{X}_i)' \ (\Sigma_i^*)^{-1} \ (X_i - \hat{X}_i)\}^{-1} \\
&\quad \left[\sum_{i=1}^{K}(X_i - \hat{X}_i)' \ (\Sigma_i^*)^{-1} \ \{\mu_i^* - E(\hat{\mu}_i^*)\}\right] \\
&= \beta + D^{-1}\sum_{i=1}^{K} M_i\{\mu_i^* - E(\hat{\mu}_i^*)\} \qquad (2.31)
\end{aligned}$$

where $D = \sum_{i=1}^{K}(X_i - \hat{X}_i)' \ (\Sigma_i^*)^{-1} \ (X_i - \hat{X}_i)$ and $M_i = (X_i - \hat{X}_i)' \ (\Sigma_i^*)^{-1}$. Using (2.21) and (2.24),

$$\begin{aligned}
E(\hat{\mu}_i^*) &= E[\hat{\gamma}(t_i) - \sum_{h=1}^{K} W_h(t_{i1}, \ldots, t_{in_i})\hat{\gamma}(t_h)] \\
&= E[\hat{y}_i - \hat{X}_i\beta] - \sum_{h=1}^{K} W_h(t_{i1}, \ldots, t_{in_i})E[\hat{y}_h - \hat{X}_h\beta] \\
&= \sum_{h=1}^{K} W_h(t_{i1}, \ldots, t_{in_i})E[y_h] - \hat{X}_i\beta \\
&\quad - \sum_{h=1}^{K} W_h(t_{i1}, \ldots, t_{in_i})E[\hat{y}_h] + \sum_{h=1}^{K} W_h(t_{i1}, \ldots, t_{in_i})\hat{X}_h\beta \\
&= \sum_{h=1}^{K} W_h(t_{i1}, \ldots, t_{in_i})[X_h\beta + \gamma(t_h)] - \hat{X}_i\beta \\
&\quad - \sum_{h=1}^{K} W_h(t_{i1}, \ldots, t_{in_i})\{\sum_{j=1}^{K} W_j(t_{h1}, \ldots, t_{hn_h})E(y_j)\} + \sum_{h=1}^{K} W_h(t_{i1}, \ldots, t_{in_i})\hat{X}_h\beta
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{h=1}^{K} W_h(t_{i1},\ldots,t_{in_i})\gamma(t_h) - \sum_{h=1}^{K} W_h(t_{i1},\ldots,t_{in_i}) \\
&\quad \{\sum_{j=1}^{K} W_j(t_{h1},\ldots,t_{hn_h})[X_j\beta+\gamma(t_j)]\} + \sum_{h=1}^{K} W_h(t_{i1},\ldots,t_{in_i})\hat{X}_h\beta \\
&= \sum_{h=1}^{K} W_h(t_{i1},\ldots,t_{in_i})\gamma(t_h) - \sum_{h=1}^{K} W_h(t_{i1},\ldots,t_{in_i})\{\sum_{j=1}^{K} W_j(t_{h1},\ldots,t_{hn_h})\gamma(t_j)\}
\end{aligned}
$$

Hence by using (2.24) we obtain

$$
\begin{aligned}
E(\hat{\mu}_i^*) &= \gamma(t_i) - \mu_i^* - \sum_{h=1}^{K} W_h(t_{i1},\ldots,t_{in_i})\{\gamma(t_h) - \mu_h^*\} \\
&= \gamma(t_i) - \mu_i^* - [\gamma(t_i) - \mu_i^*] + \sum_{h=1}^{K} W_h(t_{i1},\ldots,t_{in_i})\mu_h^* \\
&= \sum_{h=1}^{K} W_h(t_{i1},\ldots,t_{in_i})\mu_h^* = g_1(\mu_1^*,\ldots,\mu_K^*).
\end{aligned}
$$

Substituting in (2.31), the bias in estimating β amounts to

$$
\begin{aligned}
|E(\hat{\beta}_{FSSGQL}) - \beta| &= D^{-1}\sum_{i=1}^{K} M_i\{\mu_i^* - \sum_{h=1}^{K} W_h(t_{i1},\ldots,t_{in_i})\mu_h^*\} \\
&= g_2(\mu_1^*,\ldots,\mu_K^*) \qquad (2.32)
\end{aligned}
$$

where D and M_i are defined in (2.31). Note that the bias quantity in (2.32) may be negligible, provided the kernel weights are chosen satisfying (2.25). This is because under the limiting condition (2.25), $\mu_i^* \to 0$, yields bias $\to 0$.

Variance of $\hat{\beta}_{FSSGQL}$:

We now compute the variance of $\hat{\beta}_{FSSGQL}$ as

$$
\begin{aligned}
var(\hat{\beta}_{FSSGQL}) &= \{\sum_{i=1}^{K}(X_i - \hat{X}_i)'\,(\Sigma_i^*)^{-1}\,(X_i - \hat{X}_i)\}^{-1} \\
&\quad \times \sum_{i=1}^{K}(X_i - \hat{X}_i)'\,(\Sigma_i^*)^{-1}\,\{var(y_i - \hat{y}_i - \hat{\mu}_i^*)\}\,(\Sigma_i^*)^{-1}(X_i - \hat{X}_i) \\
&\quad \{\sum_{i=1}^{K}(X_i - \hat{X}_i)'\,(\Sigma_i^*)^{-1}\,(X_i - \hat{X}_i)\}^{-1} \qquad (2.33)
\end{aligned}
$$

where

$$var(y_i - \hat{y}_i - \hat{\mu}_i^*) = var(y_i - \hat{y}_i) + var(\hat{\mu}_i^*) - 2cov(y_i - \hat{y}_i, \hat{\mu}_i^*).$$

Now,

$$\begin{aligned}
var(\hat{\mu}_i^*) &= E(\hat{\mu}_i^* - E(\hat{\mu}_i^*))(\hat{\mu}_i^* - E(\hat{\mu}_i^*))' \\
&= E(\hat{\mu}_i^* - g_1(\mu_1^*, \ldots, \mu_K^*))(\hat{\mu}_i^* - g_1(\mu_1^*, \ldots, \mu_K^*))' \\
&= E\{(m_{i2}(y, x, \beta) - g_1(\mu_1^*, \ldots, \mu_K^*))(m_{i2}(y, x, \beta) - g_1(\mu_1^*, \ldots, \mu_K^*))'\} \\
&= E(m_{i2}(y, x, \beta)m_{i2}'(y, x, \beta)) - E(m_{i2}(y, x, \beta)g_1'(\mu_1^*, \ldots, \mu_K^*)) \\
&= q_1(\Sigma_i, \beta) - g_1(\mu_1^*, \ldots, \mu_K^*)g_1'(\mu_1^*, \ldots, \mu_K^*) \\
&= q_2(\Sigma_i, \beta), \text{ say}
\end{aligned}$$

and

$$\begin{aligned}
cov(y_i - \hat{y}_i, \hat{\mu}_i^*) &= cov[y_i - m_{i1}(y), m_{i2}(y, x, \beta)] \\
&= q_3(\Sigma_i, \beta) - q_4(\Sigma_i, \beta) \\
&= q_5(\Sigma_i, \beta), \text{ say}
\end{aligned}$$

where $m_{i1}(.)$ is a function of y and $m_{i2}(.)$ is a function of (y, x, β). This gives

$$var(y_i - \hat{y}_i - \hat{\mu}_i^*) = \Sigma_i^* + q_2(\Sigma_i, \beta) - 2q_5(\Sigma_i, \beta) \tag{2.34}$$

Substituting (2.34) in (2.33) we obtain

$$\begin{aligned}
var(\hat{\beta}_{FSSGQL}) &= [\sum_{i=1}^{K}(X_i - \hat{X}_i)' \ (\Sigma_i^*)^{-1} \ (X_i - \hat{X}_i)]^{-1} + \\
&\quad D^{-1}\left[\sum_{i=1}^{K}(X_i - \hat{X}_i)' \ (\Sigma_i^*)^{-1} \ \{q_2(\Sigma_i, \beta) - 2q_5(\Sigma_i, \beta)\} \ (\Sigma_i^*)^{-1}(X_i - \hat{X}_i)\right] D^{-1} \\
&= D^{-1} + L(\Sigma_1, \ldots, \Sigma_K, X_1, \ldots, X_K), \text{ say.}
\end{aligned} \tag{2.35}$$

Asymptotic distribution of $\hat{\beta}_{FSSGQL}$:

Using (2.31) and (2.35), and applying Lindeberg-Feller central limit theorem [Amemiya (1985), Theorem 3.3.6, p.92] it then follows that

$$\hat{\beta}_{FSSGQL} \sim N_p(\beta + g_2(.), D^{-1} + L(.)) \tag{2.36}$$

For convenience in our notation, the Lindeberg-Feller central limit theorem is stated as follows.

Let $y_i^* = D^{-1}\ (X_i - \hat{X}_i)'(\Sigma_i^*)^{-1}\ (y_i - \hat{y}_i - \hat{\mu}_i^*)$ so that $y_1^*, \ldots y_K^*$ are independent. Also let, $\bar{y}^* = \frac{1}{K}\sum_{i=1}^K y_i^*$ with

$$\begin{aligned} E(\bar{Y}^*) &= \frac{1}{K}[\beta + g_2(\mu_1^*, \ldots, \mu_K^*)] \\ var(\bar{Y}^*) &= \frac{1}{K^2}[D^{-1} + L(\Sigma_1, \ldots, \Sigma_K, X_1, \ldots, X_K)]. \end{aligned}$$

Then using the Lindeberg-Feller central limit theorem,

$$\bar{z}^* = (var(\bar{Y}^*))^{-1/2}(\bar{Y}^* - E(\bar{Y}^*)) \sim N_p(0, I).$$

This shows that $\hat{\beta}_{FSSGQL}$ has the p-dimensional normal distribution as stated in (2.36).

2.2.3 Estimation of ρ and σ^2

For $n = \max_{1 \le i \le K} n_i$, and

$$\delta_{iu} = \begin{cases} 1, & \text{if } u \le n_i \\ 0, & \text{if } n_i < u \le n, \end{cases}$$

the auto-correlation matrix $C_i(\rho)$ (see also (2.4)) is estimated by using the estimates of lag correlation ρ_ℓ given by

$$\hat{\rho}_\ell = \frac{\sum_{i=1}^K \sum_{u=1}^{n-\ell} \delta_{iu}\delta_{i,u+\ell}\tilde{y}_{iu}\tilde{y}_{i,u+l} / \sum_{i=1}^K \sum_{u=1}^{n-\ell} \delta_{iu}\delta_{i,u+\ell}}{\sum_{i=1}^K \sum_{u=1}^{n_i} \delta_{iu}\tilde{y}_{iu}^2 / \sum_{i=1}^K \sum_{u=1}^{n_i} \delta_{iu}}, \ \ell = 1, 2, \ldots, n-1 \tag{2.37}$$

[Sutradhar (2011, Section 2.2.2)] with $\tilde{y}_{iu} = \frac{y_{iu} - x'_{iu}\hat{\beta} - \hat{\gamma}(t_{iu})}{\hat{\sigma}}$, where $\hat{\beta}$ and $\hat{\gamma}(t)$ are the FSSGQL estimates of β and $\gamma(t)$, respectively.

The variance parameter σ^2 for the A_i matrix is estimated as

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{K}\sum_{j=1}^{n_i}(y_{ij} - x'_{ij}\hat{\beta} - \hat{\gamma}(t_{ij}))^2}{\sum_{i=1}^{K} n_i}. \tag{2.38}$$

The moment estimators for lag correlations (2.37) and variance component (2.38) are primarily developed by assuming that β and $\gamma(\cdot)$ are known, but the estimates are obtained by using the consistent estimates $\hat{\beta}_{FSSGQL}$ for β from (2.29), and $\hat{\gamma}(\cdot)$ for $\gamma(.)$ from (2.21). This leads to the consistent estimation of ρ_ℓ and σ^2 under some mild regularity conditions [Casella and Berger (1990)].

For convenience of application of the proposed FSSGQL approach, we now summarize this approach in the following four steps.

Step F1. For an initial value of β, we solve the 'working' independence assumption based semi-parametric equation (2.20) to estimate the non-parametric function $\gamma(\cdot)$.

Step F2. The estimate of $\gamma(\cdot)$ from **Step F1** and the initial β are used in (2.38) to obtain first an initial estimate of the variance component σ^2, and then initial estimates of lag correlations by (2.37).

Step F3. In this step, the estimates of auto-correlations from **Step F2** are used to compute first the kernel weights based covariance matrix $\Sigma_i^* = \text{cov}[Y_i - \hat{Y}_i]$, which is then used in (2.29) along with the estimate of μ_i^* to obtain the FSSGQL estimate of β.

Step F4. Next, the first step estimate of β from **Step F3** is applied to **Step F1** to obtain an improved estimate for the non-parametric function $\gamma(\cdot)$.

This constitute a cycle and the iteration cycles continue until convergence.

2.3 A Simulation study

The purpose of this section is to conduct a simulation study to examine the finite sample performance of the FSSGQL and various versions of the existing PSSGEE approaches in estimating the main regression parameters as well as the nuisance non-parametric function.

2.3.1 Simulation design

Sample Size:

Some of the larger data set in the longitudinal studies, such as the asthma study which contains information from $K = 537$ children, was considered to be large. Thus in this finite sample study, we choose $K = 100$ as a small sample size. Note that the asymptotic properties of the regression estimator discussed in Section 2.2.2 depends on the size of independent individuals (K), rather than on $\sum_{i=1}^{K} n_i$ as n_i responses are correlated. In longitudinal studies it is expensive and not practical to consider n_i large. We could choose variable n_i, but for simplicity we use $n_i = 4$ for $i = 1, \ldots, K$. The time points are chosen as $t_{ij} = j$ for all $i = 1, \ldots, K$, and $j = 1, \ldots, n_i$.

Covariate Selection:

We consider $p = 2$ time dependent covariates with their values as

$$x_{ij1}(t_{ij}) = \begin{cases} \frac{1}{2} & j = 1, 2 \\ & ; \ i = 1, 2, \ldots, 50, \\ 0 & j = 3, 4 \end{cases}$$

$$x_{ij1}(t_{ij}) = \begin{cases} -\frac{1}{2} & j = 1 \\ 0 & j = 2,3 \\ \frac{1}{2} & j = 4 \end{cases} ;\ i = 51, 52, \ldots, 100.$$

$$x_{ij2}(t_{ij}) = \left\{ \frac{j-2.5}{2j} \quad j = 1,2,3,4 \right. ; i = 1, 2, \ldots, 50,$$

$$x_{ij2}(t_{ij}) = \begin{cases} 0 & j = 1,2 \\ & ;\ i = 51, 52, \ldots, 100. \\ \frac{1}{2} & j = 3,4 \end{cases}$$

For the effects of these covariates we consider $(\beta_1, \beta_2)' = (1.0, 0.5)'$. By choosing these covariates, we have attempted to accommodate covariates with different natures such as categorical covariates (x_{ij1}) and mixed (i.e., categorical and continuous) covariates (x_{ij2}) in the study. We also partitioned K into two groups to include some of the practical longitudinal studies where we have two groups such as placebo and treatment.

Nonparametric function:

We consider a quadratic as well as a harmonic function for $\gamma(t_{ij})$ given by

(i) $\gamma(t_{ij}) = 3 + 2(t_{ij} - \frac{n_i+1}{2}) + (t_{ij} - \frac{n_i+1}{2})^2;\ t_{ij} = 1, 2, 3, 4; n_i = 4$

(ii) $\gamma(t_{ij}) = \sin(2t_{ij})$

In some health care related studies it may not be possible to include all possible covariates to examine their effects on the responses. However, these variables may not be ignored in some studies. We may use an unspecified function to represent such covariates if they are time dependent. Moreover, this type of function may increase as time increases. To reflect this situation we have chosen a quadratic function. We also

consider a sine function to represent other situations in practice where this unspecified time dependent function has a periodic pattern.

True Correlation structure:

We consider three correlation structures from (2.4), (2.5) and (2.6), with selected values of parameters as indicated below.

(i) **AR(1) model:** $\phi = 0.5, 0.8;\ \sigma_a^2 = 1.0$

(ii) **MA(1) model:** $\theta = 0.1, 0.4;\ \sigma_a^2 = 1.0$

(iii) **EQC model :** $\zeta = \frac{\tilde{\sigma}^2}{\tilde{\sigma}^2+\sigma_a^2} = 0.5, 0.8;\ \sigma_a^2 = 1.0$

2.3.2 Data generation and simulation results

We use various combinations of the selected design parameters to generate y_{ij} from (2.2), for $i = 1, 2, \ldots, 100$ and $j = 1, 2, 3, 4$. The simulation was repeated 1000 times.

Under each simulation we apply the four step procedure from Section 2.2 to obtain the FSSGQL estimates of β, $\gamma(t), \sigma^2$, and $\rho(\ell)$. Note that in this approach, irrespective of the true correlation models, AR(1), MA(1), or EQC, the correlation matrix is estimated by using the estimate of the general correlation matrix $C_i(\rho)$. Moreover, this approach uses corrected weight matrix in the estimating formula (2.29) for β.

Since the 'working' correlations approach does not have any guidance for the selection of correlation model, one may choose any of the low order commonly used structure such as AR(1), MA(1), EQC, or 'working' independence models [Liang and Zeger (1986)]. Thus, if data are generated from the true AR(1) model, we examine through efficiency comparison whether one can use any of the conventional low order correlation models or use the $C_i(\rho)$, which contains all these low order

correlations and provides more efficient estimates. When data are generated using, for example, a true AR(1) correlation model with high correlation such as $\phi = 0.8$, there may not exist any corresponding correlation parameter under MA(1) 'working' correlation structure [Crowder (1995)]. For this case, the moment estimates for the MA(1) correlation parameter were always more than 0.5, the boundary value. Thus, we have used $\hat{\alpha} = 0.48$ to avoid such difficulties. In Tables 2.1, 2.3, 2.4 and 2.6 this is indicated with a question mark (?). We also use the unstructured (UNS) [see Lin and Carroll (2001), for example] correlation model as a 'working' correlation model. Further, the PSSHGEE(I) and PSSHGEE based estimates discussed in Section 2.1.2 are also computed.

To simplify tables and figures, we rename the FSSGQL estimates as semi-parametric GQL (SGQL) estimates, and similarly all PSSGEE and PSSHGEE estimates as SGEE and SHGEE estimates, respectively. The efficiency of these estimates are computed by comparing their simulation-based variance with the variance of the known correlation structure based estimates, where the known correlation structure based estimates were computed by replacing the $C_i(\rho)$ matrix in the FSSGQL approach with the true correlation such as AR(1) correlation matrix.

Because the regression parameters β_1 and β_2 are of main interest, we concentrate on the efficiency performance for these two parameters. More specifically, we display the efficiencies for their estimators under various methods for a selected correlation parameter value, in Figures 2.1 and 2.2, when $\gamma(t)$ is chosen as $3 + 2(t - \frac{4+1}{2}) + (t - \frac{4+1}{2})^2$, and in Figures 2.3 and 2.4 when $\gamma(t) = \sin(2t)$. Note that the efficiencies of a selected method is computed by comparing the variance of the estimator with the corresponding variance when the estimation is based on the true correlation structure.

For example, when the data are generated using EQC structure, the efficiency of SGEE(I) for β_1 estimation, for example, is computed by $\frac{Var(\hat{\beta}_{1.EQC(true)})}{Var(\hat{\beta}_{1.SGEE(I)})}$, which was found to be 93.68% as reported in Figure 2.1. When various methods of estimation for β_1 and β_2 are compared, all methods appear to produce unbiased and hence consistent estimates for both of the regression parameters.

It is clear from Figures 2.1 and 2.2 that the proposed SGQL approach always yields the same or more efficient estimates than the other SGEE approaches including the unstructured correlations based SGEE(UNS) approach. For example, for the estimation of β_1 (Figure 2.1), under the true AR(1) correlation structure with $\phi = 0.8$ ($\rho = 0.8$) the SGQL and SGEE(EQC) provide almost equally efficient estimate whereas the other SGEE approaches including SGEE(UNS) provide less efficient estimate. Under the true MA(1) correlation model with $\theta = 0.4$ ($\rho = 0.35$), all approaches appear to produce almost equally efficient estimate for β_1, the SGEE(UNS) being slightly inferior. Similarly under the EQC process with $\zeta = 0.8$ ($\rho = 0.8$) all SGEE approaches are less efficient than the SGQL approach. Note that SGEE(I) performs the worst among all 'working' correlation approaches. Figure 2.2 shows that for the estimation of β_2, all SGEE approaches are in general inferior to the SGQL approach, the SGEE(I) being the worst followed by SGEE(MA(1)). The efficiency performances of Figures 2.3 and 2.4 are similar to Figures 2.1 and 2.2. Thus, the SGQL approach uniformly produce the same or higher efficient estimates for both β_1 and β_2 irrespective of the true correlation structures as well as non-parametric functional forms.

The efficiency of SHGEE(I) and SHGEE approaches [Fan et al (2007), Fan and Wu (2008)] discussed in Section 2.1.2 are displayed in Tables 2.1, 2.2, and 2.3, along

with other SGEE estimates for the case when $\gamma(t) = 3 + 2(t - \frac{4+1}{2}) + (t - \frac{4+1}{2})^2$. It is clear that similar to other SGEE approaches they also produce regression estimates with larger variances as compared to the SGQL estimates. Similarly, the regression estimates with $\gamma(t) = \sin(2t)$ are given in Tables 2.4, 2.5, and 2.6 for true correlation models AR(1), MA(1) and EQC respectively. The results in these tables show the same pattern as those of Tables 2.1, 2.2, and 2.3.

Further, the estimation of $\beta = (\beta_1, \beta_2)'$ requires $\gamma(t)$ which is estimated using the semi-parametric QL (SQL) estimating equation (2.9) under all SGQL and SGEE approaches. For the bandwidth b involved in the Gaussian kernel in (2.9), we have chosen $b = \frac{1}{(4K)^{\frac{1}{5}}}$ [Pagan and Ullah (1999), p.25]. For selected values of the correlation parameter, the estimates of $\gamma(t) = 3 + 2(t - \frac{4+1}{2}) + (t - \frac{4+1}{2})^2$ for all selected values of t are shown in Figures 2.5, 2.6 and 2.7 under true AR(1), MA(1) and EQC models, respectively. We have also computed the estimates for $\gamma(t) = \sin(2t)$ under all these three true correlation models, but displayed the EQC case only in Figure 2.8 as an example. The results are similar for other cases also. It is clear from these four figures that this non-parametric function is estimated very well by the semi-parametric QL approach.

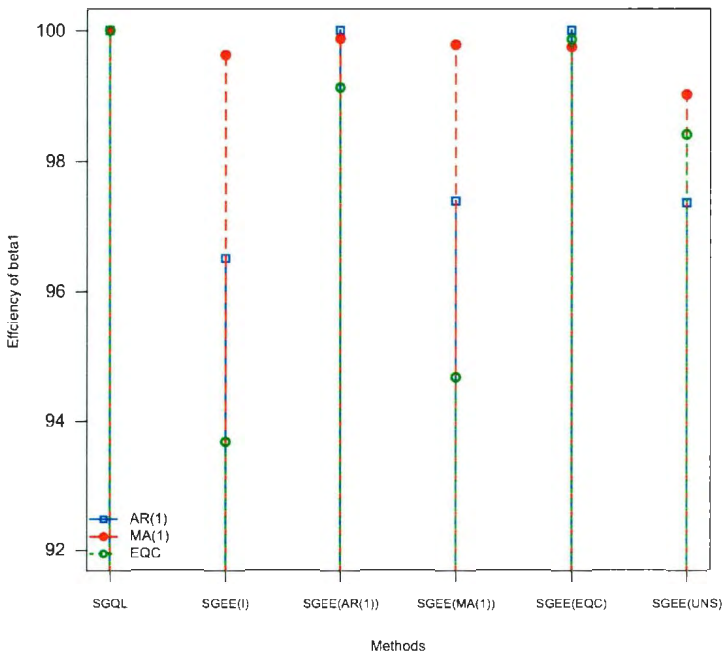

Figure 2.1: Efficiency comparisons of various semi parametric methods for the estimates of β_1 with $\gamma(t) = 3+2(t-\frac{n+1}{2})+(t-\frac{n+1}{2})^2$, under selected correlation processes: AR(1) with $\phi = 0.8$, MA(1) with $\theta = 0.4$ and EQC with $\zeta = 0.8$.

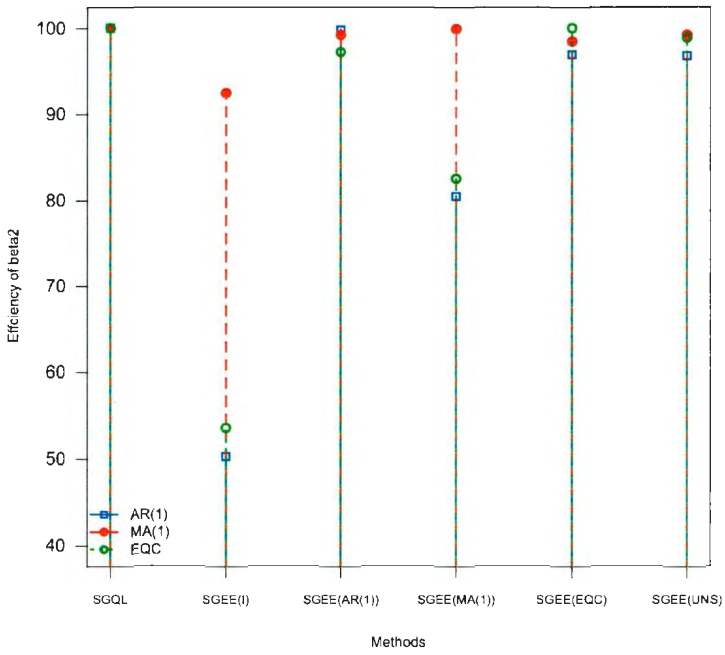

Figure 2.2: Efficiency comparisons of various semi parametric methods for the estimates of β_2 with $\gamma(t) = 3+2(t-\frac{n+1}{2})+(t-\frac{n+1}{2})^2$, under selected correlation processes: AR(1) with $\phi = 0.8$, MA(1) with $\theta = 0.4$ and EQC with $\zeta = 0.8$.

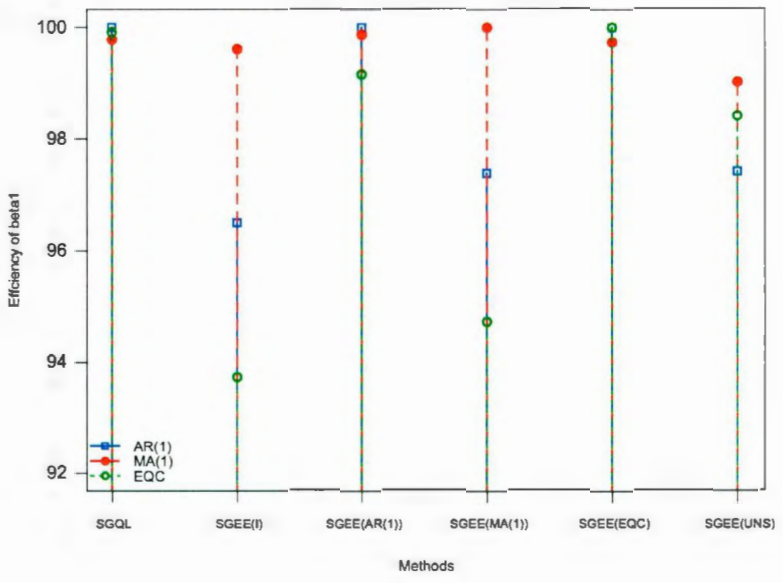

Figure 2.3: Efficiency comparisons of various semi parametric methods for the estimates of β_1 with $\gamma(t) = \sin 2t$, under selected correlation processes: AR(1) with $\phi = 0.8$, MA(1) with $\theta = 0.4$ and EQC with $\zeta = 0.8$.

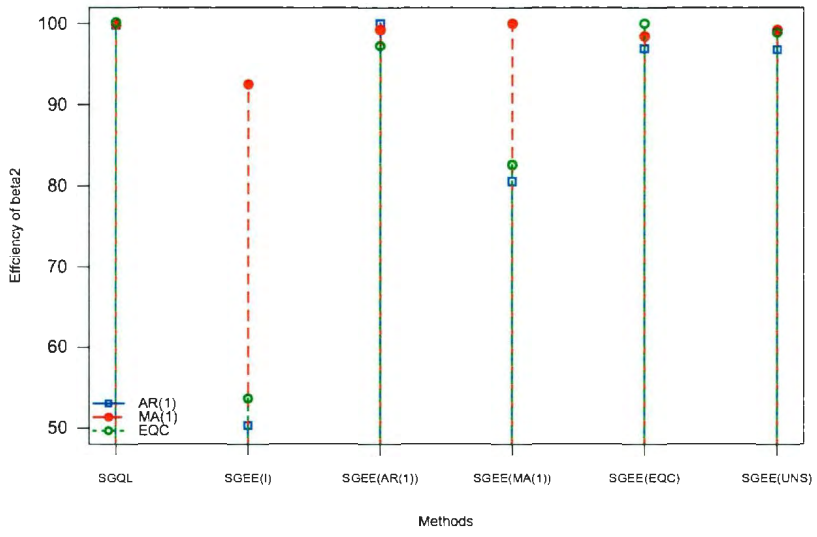

Figure 2.4: Efficiency comparisons of various semi parametric methods for the estimates of β_2 with $\gamma(t) = \sin 2t$, under selected correlation processes: AR(1) with $\phi = 0.8$, MA(1) with $\theta = 0.4$ and EQC with $\zeta = 0.8$.

Table 2.1: Simulated means (SMs) and simulated standard errors (SSEs) of the estimates of regression parameters $\beta_1 = 1$ and $\beta_2 = 0.5$, under AR(1) correlation model for selected values of the model parameters ϕ and σ^2; with $\gamma(t) = 3 + 2(t - \frac{n+1}{2}) + (t - \frac{n+1}{2})^2$; K=100; n=4; and 1000 simulations.

Estimates under the true AR(1) model

$\phi(\sigma^2)$	Method	Quantity	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\alpha}$	$\hat{\rho}_1$	$\hat{\rho}_2$	$\hat{\rho}_3$
0.5 (1.33)	SGEE(AR(1))	SM	0.9997	0.5082		0.4974		
		SSE	0.2339	0.3073		0.0580		
	SGQL	SM	0.9993	0.5072		0.4987	0.2489	0.1277
		SSE	0.2340	0.3073		0.0504	0.0728	0.0973
	SGEE(UNS)	SM	0.9999	0.5077				
		SSE	0.2365	0.3105				
	SGEE(I)	SM	0.9999	0.5094				
		SSE	0.2343	0.3715				
	SGEE(MA(1))	SM	0.9996	0.5086	0.4692			
		SSE	0.2349	0.3099	0.0251			
	SGEE(EQC)	SM	0.9998	0.5087	0.3529			
		SSE	0.2339	0.3112	0.0549			
	SHGEE(I)	SM	0.9999	0.5093				
		SSE	0.2343	0. 3722				
	SHGEE	SM	0.9991	0.5074		0.4983	0.2500	0.1292
		SSE	0.2337	0.3077		0.0477	0.0732	0.0981

Table 2.1 Continued

Estimates under the true AR(1) model

$\phi(\sigma^2)$	Method	Quantity	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\alpha}$	$\hat{\rho}_1$	$\hat{\rho}_2$	$\hat{\rho}_3$
0.8 (2.78)	SGEE(AR(1))	SM	1.0005	0.5066		0.7998		
		SSE	0.2425	0.3149		0.0298		
	SGQL	SM	1.0003	0.5057		0.8001	0.6400	0.5140
		SSE	0.2425	0.3155		0.0316	0.0504	0.0730
	SGEE(UNS)	SM	1.0013	0.5047				
		SSE	0.2491	0.3253				
	SGEE(I)	SM	1.0022	0.5181				
		SSE	0.2513	0.6259				
	SGEE(MA(1))	SM	1.0018	0.5111	0.4800(?)			
		SSE	0.2490	0.3911	-			
	SGEE(EQC)	SM	1.0011	0.5083	0.6987			
		SSE	0.2425	0.3250	0.0400			
	SHGEE(I)	SM	1.0023	0.5181				
		SSE	0.2524	0.6275				
	SHGEE	SM	1.001	0.5062		0.8001	0.6418	0.5180
		SSE	0.2450	0.3178		0.0263	0.0503	0.0735

Table 2.2: Simulated means (SMs) and simulated standard errors (SSEs) of the estimates of regression parameters $\beta_1 = 1$ and $\beta_2 = 0.5$, under MA(1) correlation model for selected values of the model parameters θ and σ^2; with $\gamma(t) = 3 + 2(t - \frac{n+1}{2}) + (t - \frac{n+1}{2})^2$; K=100; n=4; and 1000 simulations.

Estimates under the true MA(1) model

$\theta(\sigma^2)$	Method	Quantity	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\alpha}$	$\hat{\rho}_1$	$\hat{\rho}_2$	$\hat{\rho}_3$
0.1(1.01)	SGEE(MA(1))	SM	1.0014	0.4982		0.0971		
		SSE	0.2005	0.2641		0.0586		
	SGQL	SM	1.0011	0.4986		0.0971	0.0000	-0.0018
		SSE	0.2009	0.2643		0.0586	0.0684	0.0982
	SGEE(UNS)	SM	1.0001	0.4976				
		SSE	0.2026	0.2654				
	SGEE(I)	SM	1.0013	0.4992				
		SSE	0.2005	0.2650				
	SGEE(AR(1))	SM	1.0013	0.4983	0.0865			
		SSE	0.2006	0.2651	0.0810			
	SGEE(EQC)	SM	1.0012	0.4985	0.0481			
		SSE	0.2005	0.2642	0.0449			
	SHGEE(I)	SM	1.0118	0.4996				
		SSE	0.2007	0. 2656				
	SHGEE	SM	1.0015	0.4992		0.0972	-0.0000	-0.0017
		SSE	0.2014	0.2658		0.0583	0.0685	0.0990

Table 2.2 Continued

Estimates under the true MA(1) model

$\theta(\sigma^2)$	Method	Quantity	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\alpha}$	$\hat{\rho}_1$	$\hat{\rho}_2$	$\hat{\rho}_3$
0.4(1.16)	SGEE(MA(1))	SM	1.0008	0.4948		0.3435		
		SSE	0.2275	0.2781		0.0528		
	SGQL	SM	1.0004	0.4954		0.3435	-0.0007	-0.0033
		SSE	0.2280	0.2784		0.0528	0.0726	0.0973
	SGEE(UNS)	SM	0.9991	0.4949				
		SSE	0.2298	0.2802				
	SGEE(I)	SM	1.0003	0.4970				
		SSE	0.2284	0.3010				
	SGEE(AR(1))	SM	1.0004	0.4959	0.2778			
		SSE	0.2278	0.2803	0.0731			
	SGEE(EQC)	SM	1.0002	0.4962	0.1702			
		SSE	0.2281	0.2825	0.0523			
	SHGEE(I)	SM	1.0011	0.4975				
		SSE	0.2281	0. 3009				
	SHGEE	SM	1.0008	0.4969		0.3430	-0.0002	-0.0033
		SSE	0.2283	0.2784		0.0511	0.0728	0.0983

Table 2.3: Simulated means (SMs) and simulated standard errors (SSEs) of the estimates of regression parameters $\beta_1 = 1$ and $\beta_2 = 0.5$, under equi correlation model for selected values of the model parameters ζ and σ^2; with $\gamma(t) = 3+2(t-\frac{n+1}{2})+(t-\frac{n+1}{2})^2$; K=100; n=4; and 1000 simulations.

Estimates under the true EQC model

$\zeta(\sigma^2)$	Method	Quantity	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\alpha}$	$\hat{\rho}_1$	$\hat{\rho}_2$	$\hat{\rho}_3$
0.5(2.0)	SGEE(EQC)	SM	0.9967	0.5211		0.4994		
		SSE	0.2111	0.4088		0.0504		
	SGQL	SM	0.9968	0.5194		0.5003	0.4986	0.4985
		SSE	0.2115	0.4088		0.0564	0.0577	0.0870
	SGEE(UNS)	SM	0.9979	0.5205				
		SSE	0.2125	0.4118				
	SGEE(I)	SM	0.9968	0.5215				
		SSE	0.2124	0.5019				
	SGEE(AR(1))	SM	0.9967	0.5204	0.6388			
		SSE	0.2131	0.4180	0.0450			
	SGEE(MA(1))	SM	0.9969	0.5201	0.4668			
		SSE	0.2140	0.4165	0.0282			
	SHGEE(I)	SM	0.9967	0.5214				
		SSE	0.2131	0.5036				
	SHGEE	SM	0.9971	0.5195		0.5011	0.4999	0.5011
		SSE	0.2121	0.4107		0.0551	0.0579	0.0768

Table 2.3 Continued

Estimates under the true EQC model

$\zeta(\sigma^2)$	Method	Quantity	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\alpha}$	$\hat{\rho}_1$	$\hat{\rho}_2$	$\hat{\rho}_3$
0.8(5.0)	SGEE(EQC)	SM	0.9968	0.5216		0.7992		
		SSE	0.2135	0.4725		0.0274		
	SGQL	SM	0.9968	0.5192		0.7998	0.7989	0.7986
		SSE	0.2138	0.4725		0.0317	0.0296	0.0532
	SGEE(UNS)	SM	0.9983	0.5212				
		SSE	0.2170	0.4777				
	SGEE(I)	SM	0.9981	0.5325				
		SSE	0.2279	0.8811				
	SGEE(AR(1))	SM	0.9964	0.5198	0.8715			
		SSE	0.2154	0.4860	0.0188			
	SGEE(MA(1))	SM	0.9980	0.5252	0.4800(?)			
		SSE	0.2255	0.5723	-			
	SHGEE(I)	SM	0.9981	0.5325				
		SSE	0.2297	0.8828				
	SHGEE	SM	0.9975	0.5201		0.8007	0.8002	0.8010
		SSE	0.2167	0.4783		0.0288	0.0296	0.0364

Table 2.4: Simulated means (SMs) and simulated standard errors (SSEs) of the estimates of regression parameters $\beta_1 = 1$ and $\beta_2 = 0.5$, under AR(1) correlation model for selected values of the model parameters ϕ and σ^2; with $\gamma(t) = \sin 2t$; K=100; n=4; and 1000 simulations.

Estimates under the true AR(1) model

$\phi(\sigma^2)$	Method	Quantity	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\alpha}$	$\hat{\rho}_1$	$\hat{\rho}_2$	$\hat{\rho}_3$
0.5(1.33)	SGEE(AR(1))	SM	0.9995	0.5080		0.4974		
		SSE	0.2339	0.3073		0.0580		
	SGQL	SM	0.9991	0.5070		0.4987	0.2490	0.1278
		SSE	0.2340	0.3071		0.0504	0.0728	0.0973
	SGEE(UNS)	SM	0.9997	0.5074				
		SSE	0.2365	0.3105				
	SGEE(I)	SM	0.9998	0.5093				
		SSE	0.2343	0.3714				
	SGEE(MA(1))	SM	0.9993	0.5083	0.4692			
		SSE	0.2348	0.3099	0.0251			
	SGEE(EQC)	SM	0.9997	0.5085	0.3530			
		SSE	0.2333	0.3112	0.0549			

Table 2.4 Continued

Estimates under the true AR(1) model

$\phi(\sigma^2)$	Method	Quantity	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\alpha}$	$\hat{\rho}_1$	$\hat{\rho}_2$	$\hat{\rho}_3$
0.8(2.78)	SGEE(AR(1))	SM	1.0005	0.5066		0.7998		
		SSE	0.2425	0.3149		0.0298		
	SGQL	SM	1.0003	0.5057		0.8001	0.6400	0.5140
		SSE	0.2425	0.3155		0.0316	0.0504	0.0730
	SGEE(UNS)	SM	1.0013	0.5047				
		SSE	0.2491	0.3253				
	SGEE(I)	SM	1.0022	0.5181				
		SSE	0.2513	0.6259				
	SGEE(EQC)	SM	1.0011	0.5083	0.6987			
		SSE	0.2424	0.3250	0.0400			
	SGEE(MA(1))	SM	1.0018	0.5111	0.4800(?)			
		SSE	0.2490	0.3911	-			

Table 2.5: Simulated means (SMs) and simulated standard errors (SSEs) of the estimates of regression parameters $\beta_1 = 1$ and $\beta_2 = 0.5$, under MA(1) correlation model for selected values of the model parameters θ and σ^2; with $\gamma(t) = \sin 2t$; K=100; n=4; and 1000 simulations.

Estimates under the true MA(1) model

$\theta(\sigma^2)$	Method	Quantity	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\alpha}$	$\hat{\rho}_1$	$\hat{\rho}_2$	$\hat{\rho}_3$
0.1(1.01)	SGEE(MA(1))	SM	1.0014	0.4982		0.0971		
		SSE	0.2005	0.2641		0.0586		
	SGQL	SM	1.0011	0.4986		0.0971	0.0001	-0.0018
		SSE	0.2009	0.2643		0.0586	0.0684	0.0983
	SGEE(UNS)	SM	1.0001	0.4976				
		SSE	0.2026	0.2654				
	SGEE(I)	SM	1.0013	0.4992				
		SSE	0.2003	0.2650				
	SGEE(AR(1))	SM	1.0013	0.4983	0.0865			
		SSE	0.2006	0.2651	0.0810			
	SGEE(EQC)	SM	1.0012	0.4985	0.0481			
		SSE	0.2006	0.2642	0.0449			

Table 2.5 Continued

Estimates under the true MA(1) model

$\theta(\sigma^2)$	Method	Quantity	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\alpha}$	$\hat{\rho}_1$	$\hat{\rho}_2$	$\hat{\rho}_3$
0.4(1.16)	SGEE(MA(1))	SM	1.0008	0.4948		0.3435		
		SSE	0.2275	0.2781		0.0528		
	SGQL	SM	1.0004	0.4954		0.3435	-0.0006	-0.0033
		SSE	0.2280	0.2784		0.0528	0.0727	0.0973
	SGEE(UNS)	SM	0.9991	0.4949				
		SSE	0.2298	0.2802				
	SGEE(I)	SM	1.0003	0.4970				
		SSE	0.2284	0.3010				
	SGEE(AR(1))	SM	1.0004	0.49598	0.2778			
		SSE	0.2278	0.2803	0.0731			
	SGEE(EQC)	SM	1.0002	0.4962	0.1703			
		SSE	0.2281	0.2825	0.0523			

Table 2.6: Simulated means (SMs) and simulated standard errors (SSEs) of the estimates of regression parameters $\beta_1 = 1$ and $\beta_2 = 0.5$, under equi correlation model for selected values of the model parameters ζ and σ^2; with $\gamma(t) = \sin 2t$; K=100; n=4; and 1000 simulations.

Estimates under the true EQC model

$\zeta(\sigma^2)$	Method	Quantity	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\alpha}$	$\hat{\rho}_1$	$\hat{\rho}_2$	$\hat{\rho}_3$
0.5(2.0)	SGEE(EQC)	SM	0.9966	0.52087		0.4995		
		SSE	0.2111	0.4095		0.0504		
	SGQL	SM	0.9967	0.5192		0.5003	0.4987	0.4985
		SSE	0.2115	0.4088		0.0564	0.0577	0.0871
	SGEE(UNS)	SM	0.9978	0.5203				
		SSE	0.2125	0.4118				
	SGEE(I)	SM	0.9967	0.5214				
		SSE	0.2124	0.5019				
	SGEE(AR(1))	SM	0.9965	0.5202	0.6388			
		SSE	0.2131	0.4180	0.0450			
	SGEE(MA(1))	SM	0.9966	0.5198	0.4668			
		SSE	0.2139	0.4165	0.0282			

Table 2.6 Continued

Estimates under the true EQC model

$\zeta(\sigma^2)$	Method	Quantity	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\alpha}$	$\hat{\rho}_1$	$\hat{\rho}_2$	$\hat{\rho}_3$
0.8(5.0)	SGEE(EQC)	SM	0.9968	0.5216		0.7993		
		SSE	0.2135	0.4725		0.0274		
	SGQL	SM	0.9968	0.5192		0.7997	0.7989	0.7986
		SSE	0.2138	0.4719		0.0316	0.0297	0.0532
	SGEE(UNS)	SM	0.9983	0.5212				
		SSE	0.2170	0.4777				
	SGEE(I)	SM	0.9981	0.5325				
		SSE	0.2279	0.8811				
	SGEE(AR(1))	SM	0.9964	0.5198	0.8715			
		SSE	0.2154	0.4860	0.0189			
	SGEE(MA(1))	SM	0.9980	0.5252	0.4800(?)			
		SSE	0.2255	0.5723	-			

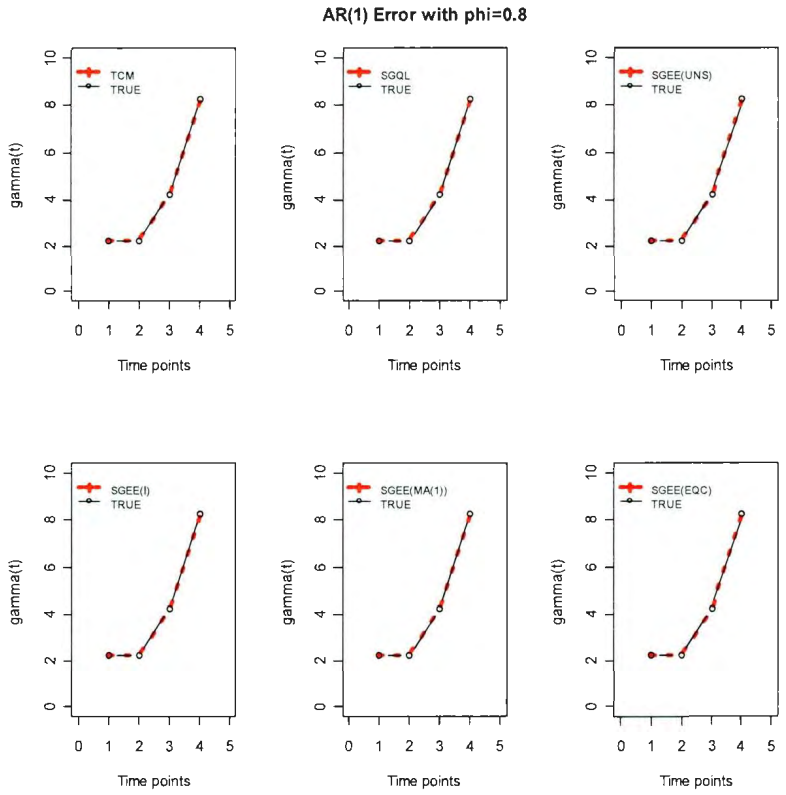

Figure 2.5: Simulated means of estimates of the non-parametric function ($\gamma(t) = 3 + 2(t - \frac{4+1}{2}) + (t - \frac{4+1}{2})^2$) under the true correlation matrix (TCM) and other selected correlation based FSSGQL method with AR(1) correlated errors.

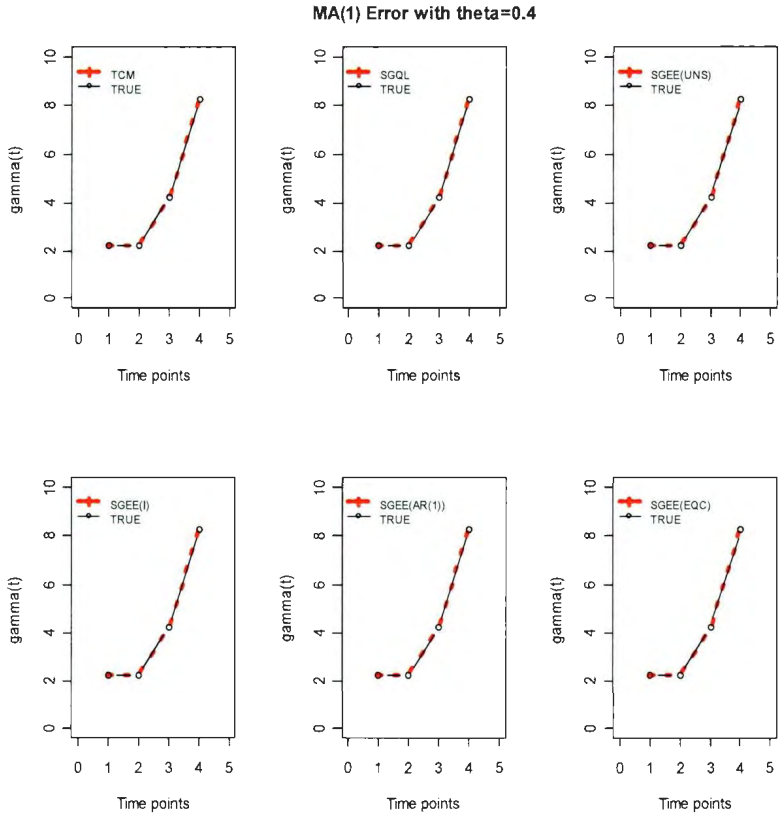

Figure 2.6: Simulated means of estimates of the non-parametric function ($\gamma(t) = 3 + 2(t - \frac{4+1}{2}) + (t - \frac{4+1}{2})^2$) under the true correlation matrix (TCM) and other selected correlation based FSSGQL method with MA(1) correlated errors.

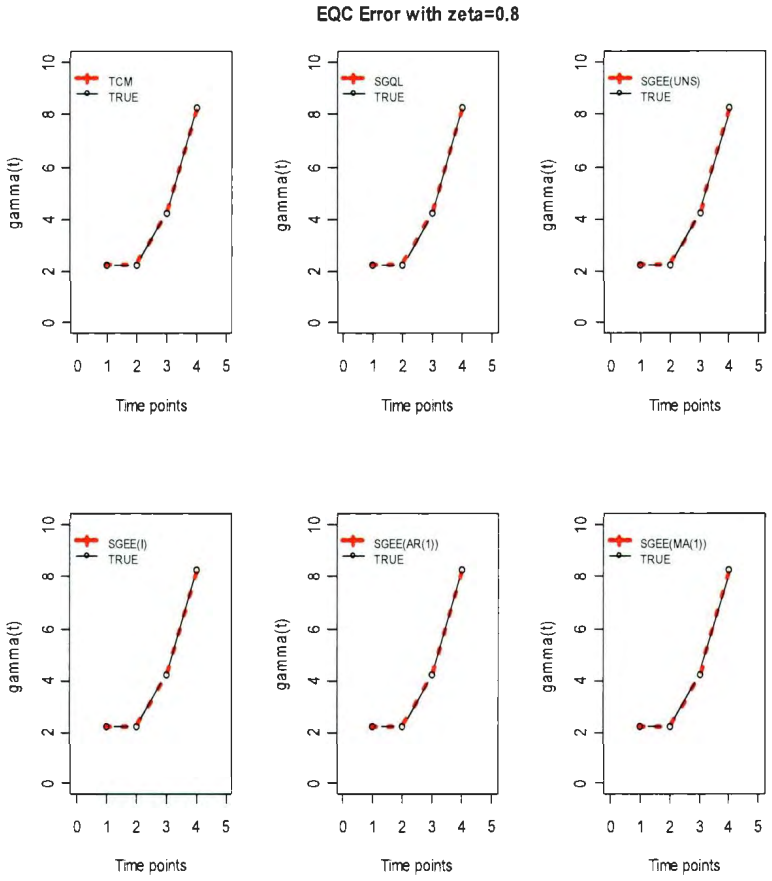

Figure 2.7: Simulated means of estimates of the non-parametric function ($\gamma(t) = 3 + 2(t - \frac{4+1}{2}) + (t - \frac{4+1}{2})^2$) under the true correlation matrix (TCM) and other selected correlation based FSSGQL method with Equi correlated errors.

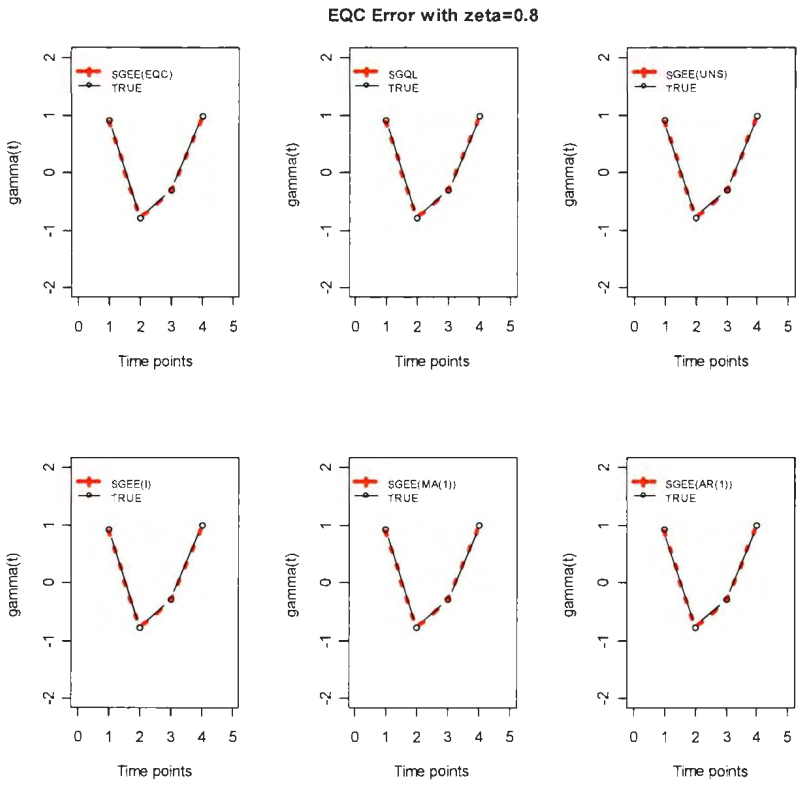

Figure 2.8: Simulated means of estimates of the non-parametric function ($\gamma(t) = \sin 2t$) under selected correlation based FSSGQL method with Equi correlated errors.

Chapter 3

Semi-parametric Longitudinal Models for Discrete Data with Non-stationary Correlation Structures

In Chapter 2, we discussed in detail the inferences for the regression effects involved in the semi-parametric linear longitudinal model. However, there are many situations where one is interested in analyzing longitudinal discrete data such as count and binary data in the semi-parametric setup. For example, we refer to the longitudinal models for the health care utilization data and Ohio asthma data mentioned in Chapter 1. But, these models do not incorporate any non-parametric functions in the regression relationship. Also, the semi-parametric inferences for linear longitudinal data discussed in Chapter 2 can not be directly generalized to the semi-parametric

longitudinal discrete data. In this chapter, we deal with such semi-parametric inferences under the assumption that the equip-spaced time based repeated responses follow a Gaussian-type ARMA class of auto-correlations. To be specific, we describe the semi-parametric longitudinal models for count data in Section 3.1 and develop the inference techniques for these models in Section 3.2. Similarly, in Sections 3.3 and 3.4, we provide the semi-parametric longitudinal models and inferences, respectively for longitudinal binary data.

3.1 Semi-parametric longitudinal models for count data with non-stationary correlation structures

Unlike the linear longitudinal models that we discussed in Chapter 2, it is traditional to consider that the count response $y_{ij}(t_{ij})$ marginally follows a Poisson distribution [Nelder and Wedderburn (1972)]. As Poisson distribution belongs to an exponential family, using the log-link function, we write the mean and variance

$$E(Y_{ij}) = V(Y_{ij}) = exp(x'_{ij}(t_{ij})\beta + \gamma(t_{ij})),$$

where $x_{ij}(t_{ij})$ is the $p-$dimensional covariate vector at time point t_{ij} and $\gamma(t_{ij})$ is the unspecified smooth function. Thus, both $x_{ij}(t_{ij})$ and $\gamma(t_{ij})$ would affect the mean response and we denote the Poisson mean and variance by

$$\mu_{ij}(\beta) = exp(x'_{ij}(t_{ij})\beta + \gamma(t_{ij})). \tag{3.1}$$

This mean function (3.1) is exactly the same as for the independent count data discussed in Section 1.2.2. However, the independence case is a special case of the

present longitudinal setup with $n_i = 1$ for all individuals $i = 1, \ldots, K$. Thus, unlike the independence setup, one has to consider the correlations of the repeated count responses $y_{i1}, \ldots, y_{ij}, \ldots, y_{in_i}$. When covariates are time dependent, the correlations in such a setup depend on these time dependent covariates which make the correlations non-stationary. We discuss this type of non-stationary correlation structures for repeated count data in the next sections.

3.1.1 Stationary correlation models for count data in semi-parametric setup

Sutradhar (2003) proposed a class of Gaussian type auto-correlation structures for stationary (time independent covariates) repeated count data which accommodates AR(1), AR(2), MA(1), ARMA(1,1), etc. and EQC correlation structures. The auto-correlation structures have the same form as $C_i(\rho)$ in (2.4). Even though the stationary correlation structures appear to be the same for linear and count data models, the dynamic relationships among the repeated responses under these models are quite different. Unlike the dynamic relationships (2.5)-(2.7) in Chapter 2, the stationary AR(1) dynamic model, for example, for count data [Sutradhar (2003)] has the form

$$y_{ij} = \rho * y_{i,j-1} + d_{ij}, \text{ for } j = 2, \ldots, n_i, \tag{3.2}$$

where $y_{i1} \sim Poi(\mu_{i.})$, $\mu_{i.} = exp(x_{i.}'(t_{i.})\beta + \gamma(t_{i.}))$, $x_{i.}(t_{i.}) = x_{ij}(t_{ij})$ and $\gamma(t_{i.}) \equiv \gamma(t_{ij})$ for all $j = 1, \ldots, n_i$. Also assume that $y_{i,j-1} \sim Poi(\mu_{i.})$ and $d_{ij} \sim Poi(\mu_{i.} - \rho\mu_{i.}), d_{ij}$ and $y_{i,j-1}$ are independent with $\rho * y_{i,j-1} = \sum_{s=1}^{y_{i,j-1}} b_s(\rho)$ with $Pr[b_s(\rho) = 1] = \rho$ and $Pr[b_s(\rho) = 0] = 1 - \rho$, ρ being the correlation parameter. This model (3.2) has the

following marginal properties:

$$\begin{aligned} E(Y_{ij}|x_{i.}) &= \mu_{i.} = exp(x'_{i.}\beta + \gamma(t_{i.})) \\ V(Y_{ij}|x_{i.}) &= \sigma_{i..} = exp(x'_{i.}\beta + \gamma(t_{i.})) \\ corr(Y_{iu}, Y_{iv}|x_{i.}, x_{i.}) &= C_i(\rho) = \rho^{|v-u|}, \text{ for all } u \neq v, |\rho| \leq 1, \end{aligned}$$

and the correlations can be represented by $C_i(\rho)$ as in (2.4), that is,

$$C_i(\rho) = \begin{pmatrix} 1 & \rho_1 & \rho_2 & \dots & \rho_{n_i-1} \\ \rho_1 & 1 & \rho_1 & \dots & \rho_{n_i-2} \\ \vdots & & & \dots & \vdots \\ \rho_{n_i-1} & \rho_{n_i-2} & & \dots & 1 \end{pmatrix} \quad \text{for all i} = 1, 2, \dots, K, \quad (3.3)$$

with $\rho_\ell = \rho^\ell$ for $\ell = 1, \dots, n_i - 1$. This $C_i(\rho)$ is also valid for stationary MA(1) and EQC correlation structures, among others. This is evident from the special cases of the non-stationary models we discussed below. These non-stationary models under the longitudinal setup for fully specified (fixed) regression functions are discussed in details in Sutradhar (2010) and (2011).

3.1.2 Non-stationary correlation models for count data

3.1.2.1 Non-stationary AR(1) models in semi-parametric setup

In the non-stationary case, the covariates are time dependent. For such cases, when the responses follow AR(1) correlation models, for example, they satisfy the dynamic relationship (3.2), that is,

$$y_{ij} = \rho * y_{i,j-1} + d_{ij}, \text{ for } j = 2, \dots, n_i. \quad (3.4)$$

However, unlike the stationary case, the marginal distributions of y_{ij} for all j are not identical. To be specific, it is now assumed that $y_{i1} \sim Poi(\mu_{i1})$, with $\mu_{i1} = exp(x'_{i1}(t_{i1})\beta + \gamma(t_{i1}))$, and for $j = 2, \ldots, n_i$, $d_{ij} \sim Poi(\mu_{ij} - \rho\mu_{i,j-1})$ where d_{ij} and $y_{i,j-1}$ are independent. Similar to the stationary case, $\rho * y_{i,j-1}$ indicates the binomial thinning operation given by $\rho * y_{i,j-1} = \sum_{s=1}^{y_{i,j-1}} b_s(\rho)$ with $Pr[b_s(\rho) = 1] = \rho$ and $Pr[b_s(\rho) = 0] = 1 - \rho$, where ρ denotes the correlation index parameter. The mean and variance of this model are given by,

$$\begin{aligned} E(Y_{ij}|x_{ij}) &= \mu_{ij} = exp(x'_{ij}(t_{ij})\beta + \gamma(t_{ij})) \\ V(Y_{ij}|x_{ij}) &= \sigma_{ijj} = \mu_{ij},\ j = 1, \ldots, n_i. \end{aligned} \tag{3.5}$$

For $j < k$, the covariance between y_{ij} and y_{ik} can be written as

$$\begin{aligned} cov(Y_{ij}, Y_{ik}|x_{ij}, x_{ik}) &= E(Y_{ij}Y_{ik}|x_{ij}, x_{ik}) - E(Y_{ij}|x_{ij})E(Y_{ik}|x_{ik}) \\ &= E_{y_{ij}} Y_{ij} E_{y_{i,k-(k-j-1)}} \cdots E_{y_{i,k-2}} E[Y_{ik}|y_{i,k-1}, y_{i,k-2}, \ldots, y_{i,k-(k-j-1)}] \\ &\quad -\mu_{ij}\ \mu_{ik} \end{aligned}$$

yielding

$$\begin{aligned} corr(Y_{ij}, Y_{ik}|x_{ij}, x_{ik}) &= c^{(ns)}_{i,j,k}(x_{ij}, x_{ik}, \rho) \\ &= \begin{cases} \rho^{k-j}\sqrt{\frac{\mu_{ij}}{\mu_{ik}}} & j < k \\ \rho^{j-k}\sqrt{\frac{\mu_{ik}}{\mu_{ij}}} & j > k \end{cases} \end{aligned} \tag{3.6}$$

with ρ satisfying the range restriction

$$0 < \rho < min[1, \frac{\mu_{ik}}{\mu_{i,k-1}}].$$

3.1.2.2 Non-stationary MA(1) models in semi-parametric setup

Suppose that $y_{i1} = d_{i1} \sim Poi(\mu_{i1}),\ i = 1, \ldots, K$, where $\mu_{i1} = exp(x'_{i1}(t_{i1})\beta + \gamma(t_{i1}))$. For the non-stationary MA(1) model, the dynamic relationship is

$$y_{ij} = \rho * d_{i,j-1} + d_{ij},\ j = 2, \ldots, n_i \tag{3.7}$$

where $d_{ij} \sim Poi(\sum_{u=0}^{j-1}(-\rho)^u \mu_{i,u-j})$ for all $j = 2, \ldots, n_i$.

The marginal properties of the model (3.7) are given by

$$\begin{aligned}
E(Y_{i1}|x_{i1}) &= \mu_{i1} = exp(x'_{i1}(t_{i1})\beta + \gamma(t_{i1})) \\
E(Y_{ij}|x_{ij}) &= E_{d_{ij},d_{i,j-1}} E[Y_{ij}|d_{ij}, d_{i,j-1}] \\
&= E_{d_{ij},d_{i,j-1}}[d_{ij} + \rho d_{i,j-1}] \\
&= \mu_{ij},\quad j = 2, \ldots, n_i. \qquad (3.8) \\
V(Y_{i1}|x_{i1}) &= \mu_{i1}
\end{aligned}$$

$$V(Y_{ij}|x_{ij}) = \mu_{ijj} = exp(x'_{ij}(t_{ij})\beta + \gamma(t_{ij})) \tag{3.9}$$

$$\begin{aligned}
corr(Y_{ij}, Y_{ik}|x_{ij}, x_{ik}) &= \begin{cases} \frac{\rho[\sum_{j=0}^{min(j,k)-1}(-\rho)^j \mu_{i,min(j,k)-j}]}{\sqrt{\mu_{ij}\mu_{ik}}} & \text{for } |j-k| = 1 \\ 0 & \text{Otherwise} \end{cases} \qquad (3.10) \\
&= c^{(ns)}_{i,j,k}(x_{ij}, x_{ik}, \rho)
\end{aligned}$$

with ρ satisfying the range restriction

$$0 < \rho < min[1, \rho_{i20}, \ldots, \rho_{ij0}, \ldots, \rho_{in_i0}],$$

where ρ_{ij0} is the solution of $\sum_{u=0}^{j-1}(-\rho)^u \mu_{i,j-u} = 0$.

But in the stationary correlation case the correlation in (3.10) reduces to

$$corr(y_{ij}, y_{ik}) = \begin{cases} \rho\{\sum_{j=o}^{\infty}(-\rho)^j = \frac{\rho}{1+\rho}\} & \text{for } |j-k| = 1 \\ \\ 0 & \text{Otherwise,} \end{cases}$$

which satisfies the correlation structure $C_i(\rho)$ in (3.3).

3.1.2.3 Non-stationary EQC models in semi-parametric setup

Assume that $y_{i1} \sim Poi(\mu_{i1}),\ i = 1, \ldots, K$ and consider the dynamic relation

$$y_{ij} = \rho * y_{i1} + d_{ij},\ j = 2, \ldots, n_i, \tag{3.11}$$

where $d_{ij} \sim Poi(\mu_{ij} - \rho\mu_{i1})$ and d_{ij} for $j = 2, 3, \ldots, n_i$ are independent to y_{i1}.

The mean and variance of the model (3.11) are

$$\begin{aligned} E(Y_{ij}|x_{ij}) &= \mu_{ij} = exp(x'_{ij}(t_{ij})\beta + \gamma(t_{ij})) \\ V(Y_{ij}|x_{ij}) &= \sigma_{ijj} = exp(x'_{ij}(t_{ij})\beta + \gamma(t_{ij})) \end{aligned} \tag{3.12}$$

For $j < k$,

$$cov[(Y_{ij}, Y_{ik})|x_{i1}] = \rho\mu_{i1}$$

yielding

$$\begin{aligned} corr(Y_{ij}, Y_{ik}|x_{ij}, x_{ik}) &= \frac{\rho\mu_{i1}}{\sqrt{\mu_{ij}\mu_{ik}}} \\ &= c^{(ns)}_{i,j,k}(x_{ij}, x_{ik}, \rho) \end{aligned} \tag{3.13}$$

with ρ satisfying the range restriction

$$0 < \rho < min[1, \frac{\mu_{ik}}{\mu_{i1}}].$$

Note that for the stationary case, the above correlation becomes

$$corr(Y_{ij}, Y_{ik}|x_{ij}, x_{ik}) = \rho, \text{ for all } j, k = 1, \ldots, n_i,$$

and this satisfies the correlation structure $C_i(\rho)$ in (3.3).

3.2 Estimation in semi-parametric models for longitudinal count data

The main interest of this section is to find the effect of the covariates $x_{ij}(t_{ij})$ on the response y_{ij} where $E(Y_{ij})$ is given in (3.1). We assume that this marginal property holds for any of the non-stationary auto-correlation models such as AR(1) (3.4), MA(1) (3.7) and EQC (3.11) models for repeated responses $y_{i1}, \ldots, y_{ij}, \ldots, y_{in_i}$. As far as $\gamma(t_{ij})$ is concerned, one may treat this as a nuisance function, which is of secondary interest. Thus, while we will exploit the correlations of the repeated data in estimating the regression parameter β, we will however estimate the nuisance function by pretending that the repeated responses are independent.

3.2.1 Estimation of non-parametric function $\gamma(.)$

When correlations are ignored, we may follow the estimating equation (1.17) developed for independent count data from Chapter 1 to estimate the non-parametric function. Thus at a given time point $t_{ij} = t_0$, we now use the semi-parametric QL estimating equation for estimating $\gamma(t_{ij})|_{t_{ij}=t_0}$ given by

$$\sum_{i=1}^{K} \sum_{j=1}^{n_i} w_{ij}(t_0) \frac{\partial \mu_{ij}}{\partial \gamma(t_0)} (\frac{y_{ij} - \mu_{ij}}{\mu_{ij}}) = 0, \tag{3.14}$$

where μ_{ij} is $exp(x'_{ij}(t_{ij})\beta + \gamma(t_{ij}))$. For convenience following the formula in (1.19) for $\hat{\gamma}(.)$, we now write

$$\hat{\gamma}(t_0) \;=\; log\left(\frac{\sum_{i=1}^{K}\sum_{j=1}^{n_i} w_{ij}(t_0)y_{ij}}{\sum_{i=1}^{K}\sum_{j=1}^{n_i} w_{ij}(t_0)exp(x'_{ij}(t_{ij})\beta)}\right) \tag{3.15}$$

where $w_{ij}(t_0) = \frac{p_{ij}(\frac{t_0-t_{ij}}{b})}{\sum_{i=1}^{K}\sum_{j=1}^{n_i} p_{ij}(\frac{t_0-t_{ij}}{b})}$, $p_{ij}(\frac{t_0-t_{ij}}{b}) = \frac{1}{\sqrt{2\pi}b}exp(\frac{-1}{2}(\frac{t_0-t_{ij}}{b})^2)$, b is the bandwidth and t_{ij} is the time measure for the i^{th} individual at time point j. When β is known or estimated, one can estimate $\gamma(t_0)$ by (3.15).

3.2.2 Estimation of β

For the estimation of β, we first express the mean response as a function of $\hat{\gamma}(.)$ from (3.15) as

$$\tilde{\mu}_{ij} = exp[x'_{ij}(t_{ij})\beta + \hat{\gamma}(t_{ij}, \beta)] \tag{3.16}$$

It is clear that the consistent and efficient estimation of β requires the consistent estimation of $\gamma(.)$ such as by (3.15) as well as the use of proper correlation structure of the repeated responses.

3.2.2.1 Naive GQL estimation approach

When the non-parametric function affects the mean response as in (3.16), it is understandable that ignoring $\gamma(.)$ in the mean while estimating β would cause a biased and inconsistent estimate. This can be examined by checking the performance of the estimate of β obtained from a naive GQL estimating equation given by

$$\sum_{i=1}^{K} \frac{\partial(\mu_i^{**})'}{\partial\beta}\; [\Sigma_i^{(ns)}(\rho)]^{-1}\; (y_i - \mu_i^{**}) = 0 \tag{3.17}$$

where $\mu_i^{**} = [\mu_{i1}^{**}, \ldots, \mu_{ij}^{**} \ldots \mu_{in_i}^{**}]'$ with $\mu_{ij}^{**} = exp[x_{ij}'(t_{ij})\beta]$, $\frac{\partial(\mu_i^{**})'}{\partial\beta} = X_i'$ and $\Sigma_i^{(ns)}(\rho) = A_i^{1/2}\ C_i^{(ns)}(\hat{\rho})\ A_i^{1/2}$ with $A_i = \text{diag}[\mu_{i1}^{**}, \ldots, \mu_{ij}^{**}, \ldots \mu_{in_i}^{**}]$. It is improtant to recognize that the responses $y_i = (y_{i1}, \ldots, y_{ij}, \ldots, y_{in_i})$ are generated with $E(Y_{ij}) = \mu_{ij} = exp[x_{ij}'(t_{ij})\beta + \gamma(t_{ij})]$ for all $i = 1, \ldots, n_i$ and this leads to $[E(Y_{ij}) - \mu_{ij}^{**}] \neq 0$. To compute $\hat{\beta}_{NGQL}$, we solve (3.17) by using the Newton-Raphson method.

Next, to obtain a consistent and efficient estimate for β involved in (3.16), we accommodate the estimation effect of $\gamma(.)$ and the correlation structure of the responses to develop the appropriate estimating equations. We do this estimation in two ways.

(1) To develop the semi-parametric GQL (SGQL) estimation similar to the existing semi-parametric GEE (SGEE) approaches, we use $var(Y_i) = \Sigma_i^{(ns)}(\rho)$ as the longitudinal weight matrix in the estimating equation. Note that $\Sigma_i^{(ns)}(\rho)$ is a correct covariance matrix under the auto-correlation class but it is different than the 'working' covariance matrix $V_i(\alpha)$ used in SGEE. We will refer to this as the partially standardized SGQL (PSSGQL) approach and will discuss this in Section 3.2.2.2.

(2) We use $var(Y_i - \tilde{\mu}_i) = \Sigma_i^{*(ns)}(\rho)$ to construct the SGQL estimating equation because of the fact that $\tilde{\mu}_i$ contains $\hat{\gamma}(.)$ which is a function of y's. We refer to this as the fully standardized SGQL (FSSGQL) approach. This is discussed in Section 3.2.2.4.

3.2.2.2 PSSGQL estimation under non-stationary (ns) correlation structure

As it is significant to consider the estimation effect of $\hat{\gamma}(.)$ for the efficient estimation of β, we propose the non-stationary correlation structures based PSSGQL(ns)

estimating equation for β as

$$\sum_{i=1}^{K} \frac{\partial \tilde{\mu}_i'}{\partial \beta} \; [\Sigma_i^{(ns)}(\rho)]^{-1} \; (y_i - \tilde{\mu}_i) \;\; = \;\; 0, \tag{3.18}$$

where

$$\Sigma_i^{(ns)}(\hat{\rho}) = \hat{var}(Y_i) = A_i^{1/2} \; C_i^{(ns)}(\hat{\rho}) \; A_i^{1/2}, \tag{3.19}$$

with $A_i = \text{diag}[\tilde{\mu}_{i1}, \ldots, \tilde{\mu}_{ij}, \ldots, \tilde{\mu}_{in_i}]$ and $C_i^{(ns)}(\hat{\rho})$ is the estimate of the $n_i \times n_i$ non-stationary correlation matrix $C_i^{(ns)}(\rho)$ defined as

$$C_i^{(ns)}(\rho) = (c_{i,j,k}^{(ns)}(x_{ij}, x_{ik}, \rho)). \tag{3.20}$$

The formulas for the elements $c_{i,j,k}^{(ns)}(.)$ depends on the correlation structures discussed in Section 3.1.2.

By using (3.16) the elements in the gradient functions are calculated as

$$\begin{aligned} \frac{\partial \tilde{\mu}_{ij}'}{\partial \beta} &= \frac{\partial}{\partial \beta} \left(exp[x_{ij}'(t_{ij})\beta + \hat{\gamma}(t_{ij}, \beta)] \right) \\ &= exp[x_{ij}'(t_{ij})\beta + \hat{\gamma}(t_{ij}, \beta)] \left(x_{ij}(t_{ij}) + \frac{\partial}{\partial \beta} \hat{\gamma}(t_{ij}) \right) \\ &= \tilde{\mu}_{ij} \left[x_{ij}(t_{ij}) - \frac{\sum_{l=1}^{K} \sum_{u=1}^{n_l} w_{lu}(t_{ij}) exp(x_{lu}'(t_{lu})\beta) x_{lu}(t_{lu})}{\sum_{l=1}^{K} \sum_{u=1}^{n_l} w_{lu}(t_{ij}) exp(x_{lu}'(t_{lu})\beta)} \right]. \end{aligned}$$

Clearly, the estimating equation in (3.18) incorporates the non-stationary correlations among the repeated count responses, which will enhance the efficiency of the regression parameter estimate. Since there does not exist any closed-form expression for β, we use Newton-Raphson method to solve (3.18).

Let

$$f(\beta) = \sum_{i=1}^{K} \frac{\partial \tilde{\mu}_i'}{\partial \beta} \; [\Sigma_i^{(ns)}(\rho)]^{-1} \; (y_i - \tilde{\mu}_i).$$

Starting with an initial value for β, each step of the following iterative equation

$$\hat{\beta}_{(r+1)} = \hat{\beta}_{(r)} - [(f'(\beta))^{-1} f(\beta)]_{\beta=\beta_{(r)}} \tag{3.21}$$

updates the value of β until convergence. The derivative function, $f'(\beta)$ at $\beta = \beta_{(r)}$ in (3.21) is calculated as

$$f'(\beta) \;=\; -\sum_{i=1}^{K} \frac{\partial \tilde{\mu}_i'}{\partial \beta} \, [\Sigma_i^{(ns)}]^{-1} \, \frac{\partial \tilde{\mu}_i}{\partial \beta'}.$$

The estimation of non-stationary correlations are slightly different than the stationary case as it subsumes the time dependent covariates in their estimation. As it is necessary to incorporate this difference, the correlation matrix $C_i^{(ns)}(\rho)$ in (3.19) has the form (3.6), (3.10) and (3.13) under the non-stationary AR(1), MA(1) and EQC correlation structures, respectively, for the estimation of β using PSSGQL(ns) approach.

Note that solving the estimating equation (3.18) requires the estimation of ρ parameter involved in the $C_i^{(ns)}(\rho)$ matrix. This correlation index parameter can be estimated consistently by using the well-known method of moments. However, the formula for ρ estimate will be different under various non-stationary correlation structures. For example, in the next section we provide the estimate of ρ under non-stationary AR(1) correlation structure $C_i^{(ns)}(\rho)$. The estimate of ρ under other non-stationary correlation structures may be obtained similarly.

3.2.2.3 Estimation of correlation index parameter ρ

In order to use the method of moment technique to estimate the correlation index parameter ρ, one can equate the sample covariance with its population counterpart

as

$$\begin{aligned}
\frac{\sum_{i=1}^{K}\sum_{j=2}^{n_i} y_{ij}^{*}y_{i,j-1}^{*}}{\sum_{i=1}^{K}(n_i-1)} &= E\left[\frac{\sum_{i=1}^{K}\sum_{j=2}^{n_i} y_{ij}^{*}y_{i,j-1}^{*}}{\sum_{i=1}^{K}(n_i-1)}\right], \quad \text{where } y_{ij}^{*}=\frac{y_{ij}-\mu_{ij}}{\sqrt{\mu_{ij}}} \\
&= \frac{1}{\sum_{i=1}^{K}(n_i-1)}E\left[\sum_{i=1}^{K}\sum_{j=2}^{n_i}\frac{(y_{ij}-\mu_{ij})}{\sqrt{\mu_{ij}}}\frac{(y_{i,j-1}-\mu_{i,j-1})}{\sqrt{\mu_{i,j-1}}}\right] \\
&= \frac{1}{\sum_{i=1}^{K}(n_i-1)}\sum_{i=1}^{K}\sum_{j=2}^{n_i}\frac{corr(Y_{ij},Y_{i,j-1})\sqrt{\mu_{ij}}\sqrt{\mu_{i,j-1}}}{\sqrt{\mu_{ij}}\sqrt{\mu_{i,j-1}}} \\
&= \frac{1}{\sum_{i=1}^{K}(n_i-1)}\sum_{i=1}^{K}\sum_{j=2}^{n_i}\rho\,\frac{\sqrt{\mu_{i,j-1}}}{\sqrt{\mu_{ij}}}, \quad \text{from (3.6)}
\end{aligned}$$

However, for the estimation of ρ, following Sutradhar (2010) we use sample auto-correlations and equate that to its population counterpart. That is,

$$\begin{aligned}
\frac{\frac{\sum_{i=1}^{K}\sum_{j=2}^{n_i} y_{ij}^{*}y_{i,j-1}^{*}}{\sum_{i=1}^{K}(n_i-1)}}{\frac{\sum_{i=1}^{K}\sum_{j=1}^{n_i}(y_{ij}^{*})^2}{\sum_{i=1}^{K}n_i}} &\approx \frac{E\left[\frac{\sum_{i=1}^{K}\sum_{j=2}^{n_i} y_{ij}^{*}y_{i,j-1}^{*}}{\sum_{i=1}^{K}(n_i-1)}\right]}{E\left[\frac{\sum_{i=1}^{K}\sum_{j=1}^{n_i}(y_{ij}^{*})^2}{\sum_{i=1}^{K}n_i}\right]} \\
&= \frac{1}{\sum_{i=1}^{K}(n_i-1)}\sum_{i=1}^{K}\sum_{j=2}^{n_i}\rho\,\frac{\sqrt{\mu_{i,j-1}}}{\sqrt{\mu_{ij}}},
\end{aligned}$$

yielding

$$\hat{\rho}=\frac{\sum_{i=1}^{K}\sum_{j=2}^{n_i} y_{ij}^{*}y_{i,j-1}^{*}}{\sum_{i=1}^{K}\sum_{j=1}^{n_i} y_{ij}^{*2}}\;\frac{\sum_{i=1}^{K}n_i}{\sum_{i=1}^{K}\sum_{j=2}^{n_i}[\frac{\hat{\mu}_{i,j-1}}{\hat{\mu}_{ij}}]^{\frac{1}{2}}}, \tag{3.22}$$

under the non-stationary AR(1) correlation model, where $y_{ij}^{*}=\frac{y_{ij}-\hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}}}$ with $\hat{\mu}_{ij}=exp(x_{ij}'(t_{ij})\hat{\beta}+\hat{\gamma}(t_{ij}))$.

PSSGQL estimation under stationary (s) correlation structure

For convenience, we refer to the PSSGQL estimation approach to PSSGQL(s) under the stationary(s) correlation structure. In this approach, we estimate the regression parameter β using a similar estimating equation as that of the non-stationary

case (3.18). The difference between the two cases lies in the fact that we now use

$$\hat{var}(Y_i) = \Sigma_i^{(s)}(\hat{\rho}) = A_i^{1/2}\ C_i(\hat{\rho})\ A_i^{1/2}, \tag{3.23}$$

with $C_i(\rho)$ as in (3.3), whereas in (3.18), the variance estimate is $\hat{var}(Y_i) = \Sigma_i^{(ns)}(\hat{\rho})$. Thus, in the present stationary case, the estimating equation has the form

$$\sum_{i=1}^{K} \frac{\partial \tilde{\mu}_i'}{\partial \beta}\ [\Sigma_i^{(s)}(\rho)]^{-1}\ (y_i - \tilde{\mu}_i) = 0.$$

Because the computation of the $\Sigma_i^{(s)}(\rho)$ requires the calculation of auto-correlation matrix $C_i(\rho)$, we estimate the lag correlations ρ_ℓ $(\ell = 1, \ldots, n_i - 1)$ as

$$\hat{\rho}_\ell = \frac{\sum_{i=1}^{K}\sum_{u=1}^{n-\ell} \delta_{iu}\delta_{i,u+\ell} y^*_{iu} y^*_{i,u+l} / \sum_{i=1}^{K}\sum_{u=1}^{n-\ell} \delta_{iu}\delta_{i,u+\ell}}{\sum_{i=1}^{K}\sum_{u=1}^{n_i} \delta_{iu} y^{*2}_{iu} / \sum_{i=1}^{K}\sum_{u=1}^{n_i} \delta_{iu}},\ \ell = 1, 2, \ldots, n-1 \tag{3.24}$$

where $n = \max_{1 \le i \le K} n_i$, and

$$\delta_{iu} = \begin{cases} 1, & \text{if } u \le n_i \\ 0, & \text{if } n_i < u \le n, \end{cases}$$

with $y^*_{iu} = \frac{y_{iu} - exp(x'_{iu}\hat{\beta} + \hat{\gamma}(t_{iu}))}{\sqrt{exp(x'_{iu}\hat{\beta} + \hat{\gamma}(t_{iu}))}}$. This formula for $\hat{\rho}_\ell$ is the same as (2.37) in Chapter 2 for linear correlated models except that an appropriate mean and variance for count data is used in the present formula. Note that the $C_i(\rho)$ matrix in (3.23) (see also (3.3)) holds for a general class of auto-correlation structures, whereas $C_i^{(ns)}(\rho)$ matrix under the non-stationary correlation models are estimated for specified correlation structures.

3.2.2.4 FSSGQL estimation under non-stationary correlation structure

The proposed PSSGQL(ns) estimating equation is constructed by using the true non-stationary covariance matrix $var(Y_i)$ as the longitudinal weights. However, as argued

in Chapter 2 under semi-parametric linear models, it is appropriate to use the weight matrix $var(Y_i - \tilde{\mu}_i)$ to construct the estimating equation for β. This adjustment arises mainly because the non-parametric function (when estimated) involved in the semi-parametric model depends on β. Also, when β is unknown, $\hat{\gamma}(t_0)$ by (3.15) still contains $\{y_{ij}\}$. Because of this reason, one should consider $\frac{\partial E(\tilde{\mu}_i')}{\partial \beta}$ as the correct gradient function while constructing the estimating equation for β. Hence, similar to the FSSGQL estimation method discussed in Chapter 2 (Section 2.2) for linear models, for $\tilde{\mu}_i = [\tilde{\mu}_{i1}, \ldots, \tilde{\mu}_{ij}, \ldots \tilde{\mu}_{in_i}]'$, one may use the FSSGQL(ns) estimating equation

$$\sum_{i=1}^{K} \frac{\partial E(\tilde{\mu}_i')}{\partial \beta} \; [var(Y_i - \tilde{\mu}_i)]^{-1} \; (y_i - \tilde{\mu}_i) = 0 \tag{3.25}$$

where $\tilde{\mu}_{ij}$ is given by (3.16) and $var(Y_i - \tilde{\mu}_i) = \Sigma_i^{*(ns)}(\rho)$, for estimating β. We use the formula for $\hat{\gamma}(t_0)$ from (3.15) and write

$$\begin{aligned} \frac{\partial E(\tilde{\mu}_{ij})}{\partial \beta} &= \frac{\partial}{\partial \beta}\left[exp(x_{ij}'(t_{ij})\beta) \; E_y \left(\frac{\sum_{l=1}^{K}\sum_{u=1}^{n_l} w_{lu}(t_{ij}) y_{lu}}{\sum_{l=1}^{K}\sum_{u=1}^{n_l} w_{lu}(t_{ij}) exp(x_{lu}'(t_{lu})\beta)} \right) \right] \\ &= \frac{\partial}{\partial \beta}\left[exp(x_{ij}'(t_{ij})\beta) \frac{\sum_{l=1}^{K}\sum_{u=1}^{n_l} w_{lu}(t_{ij}) exp(x_{lu}'(t_{lu})\beta + \gamma(t_{lu}))}{\sum_{l=1}^{K}\sum_{u=1}^{n_l} w_{lu}(t_{ij}) exp(x_{lu}'(t_{lu})\beta)} \right] \end{aligned}$$

Let $w_{lu}^*(t_{ij}) = \frac{w_{lu}(t_{ij}) exp(x_{lu}'(t_{lu})\beta}{\sum_{l=1}^{K}\sum_{u=1}^{n_l} w_{lu}(t_{ij}) exp(x_{lu}'(t_{lu})\beta)}$, then

$$\begin{aligned} \frac{\partial E(\tilde{\mu}_{ij})}{\partial \beta} &= \frac{\partial}{\partial \beta}\left[exp(x_{ij}'(t_{ij})\beta) \sum_{l=1}^{K}\sum_{u=1}^{n_l} w_{lu}^*(t_{ij}) exp(\gamma(t_{lu})) \right] \\ &= \left[\sum_{l=1}^{K}\sum_{u=1}^{n_l} w_{lu}^*(t_{ij}) exp(\gamma(t_{lu})) \right] \; exp(x_{ij}'(t_{ij})\beta) \; x_{ij}(t_{ij}). \end{aligned} \tag{3.26}$$

In order to construct the FSSGQL(ns) estimating equation (3.25), we now provide the formula to compute $\Sigma_i^{*(ns)}(\rho) = var(Y_i - \tilde{\mu}_i)$ under the present semi-parametric model for count data. However, to obtain solutions for such FSSGQL(ns) estimating equation will naturally be complicated numerically.

Computation of $\Sigma_i^{*(ns)}(\rho) = var(Y_i - \tilde{\mu}_i)$

$$\begin{aligned}\Sigma_i^{*(ns)} &= var(Y_i - \tilde{\mu}_i) \\ &= Cov(Y_i) + Cov(\tilde{\mu}_i) - 2\ Cov(Y_i, \tilde{\mu}_i') \\ &= \Sigma_i^{(ns)} + \tilde{\Sigma}_i^{(ns)} - 2\ \tilde{\tilde{\Sigma}}_i^{(ns)} \end{aligned} \tag{3.27}$$

where $\Sigma_i^{(ns)} = var(Y_i)$ has the form given in (3.19). The formulas for the calculation of the elements in $\tilde{\Sigma}_i^{(ns)}$ and $\tilde{\tilde{\Sigma}}_i^{(ns)}$ are as follows.

Computation of $\tilde{\Sigma}_i^{(ns)}$:

Because $\tilde{\mu}_i = [\tilde{\mu}_{i1}, \ldots, \tilde{\mu}_{ij}, \ldots, \tilde{\mu}_{in_i}]'$, we need to compute the the elements $var(\tilde{\mu}_{ij})$ and $cov(\tilde{\mu}_{ij}, \tilde{\mu}_{ik})$ to construct $\tilde{\Sigma}_i^{(ns)}$ matrix. The derivation for these components are given below.

First,

$$\begin{aligned} V(\tilde{\mu}_{ij}) &= V\left(exp(x_{ij}'(t_{ij})\beta + \hat{\gamma}(t_{ij}))\right) \\ &= V\left(exp(x_{ij}'(t_{ij})\beta)\frac{\sum_{l=1}^{K}\sum_{u=1}^{n_l} w_{lu}(t_{ij})y_{lu}}{\sum_{l=1}^{K}\sum_{u=1}^{n_l} w_{lu}(t_{ij})exp(x_{lu}'\beta)}\right) \\ &= \frac{exp(2x_{ij}'(t_{ij})\beta)}{(\sum_{l=1}^{K}\sum_{u=1}^{n_l} w_{lu}(t_{ij})exp(x_{lu}'\beta))^2}\sum_{l=1}^{K}\sum_{u=1}^{n_l}\sum_{v=1}^{n_l} w_{lu}(t_{ij})w_{lv}(t_{ij})\sigma_{luv}^{(ns)} \end{aligned}$$

Next,

$$\begin{aligned} Cov(\tilde{\mu}_{ij}, \tilde{\mu}_{ik}) &= Cov\left(exp(x'_{ij}(t_{ij})\beta + \hat{\gamma}(t_{ij})),\ exp(x'_{ik}(t_{ik})\beta + \hat{\gamma}(t_{ik}))\right) \\ &= exp[(x_{ij}(t_{ij}) + x_{ik}(t_{ik}))'\beta] \\ &\quad Cov\left(\frac{\sum_{l=1}^{K}\sum_{u=1}^{n_l} w_{lu}(t_{ij})y_{lu}}{\sum_{l=1}^{K}\sum_{u=1}^{n_l} w_{lu}(t_{ij})exp(x'_{lu}\beta)},\ \frac{\sum_{m=1}^{K}\sum_{v=1}^{n_m} w_{mv}(t_{ik})y_{mv}}{\sum_{m=1}^{K}\sum_{v=1}^{n_m} w_{mv}(t_{ik})exp(x'_{mv}\beta)}\right) \\ &= \frac{exp[(x_{ij} + x_{ik})'\beta]}{[\sum_{l=1}^{K}\sum_{u=1}^{n_l} w_{lu}(t_{ij})exp(x'_{lu}\beta)]\ [\sum_{m=1}^{K}\sum_{v=1}^{n_m} w_{mv}(t_{ik})exp(x'_{mv}\beta)} \\ &\quad Cov\left(\sum_{l=1}^{K}\sum_{u=1}^{n_l} w_{lu}(t_{ij})y_{lu},\ \sum_{m=1}^{K}\sum_{v=1}^{n_m} w_{mv}(t_{ik})y_{mv}\right) \end{aligned}$$

Since y_{lu}'s are independent, $Cov(y_{lu}, y_{mv}) = 0$ for all $l \neq m$ and $u \neq v$. Hence

$$Cov(\tilde{\mu}_{ij}, \tilde{\mu}_{ik}) = \frac{exp[(x_{ij}(t_{ij}) + x_{ik}(t_{ik}))'\beta]\ \sum_{l=1}^{K}\sum_{u=1}^{n_l}\sum_{v=1}^{n_l} w_{lu}(t_{ij})\ w_{lv}(t_{ik})\sigma^{(ns)}_{luv}}{[\sum_{l=1}^{K}\sum_{u=1}^{n_l} w_{lu}(t_{ij})exp(x'_{lu}\beta)]\ [\sum_{l=1}^{K}\sum_{u=1}^{n_l} w_{lu}(t_{ik})exp(x'_{lu}\beta)]}.$$

Note that when the correlations are stationary, $\sigma^{(ns)}_{luu}$ is replaced by $\sigma^{(s)}_{luu} = \mu_{lu}$ and $\sigma^{(ns)}_{luv}$ is replaced by $\sigma^{(s)}_{luv} = \rho^{v-u}\sigma_{iuu}$ for $u < v$.

Computation of $\tilde{\tilde{\Sigma}}^{(ns)}_i$:

The calculations of $\tilde{\tilde{\Sigma}}^{(ns)}_i$ matrix involves the calculation of $Cov(y_{ij}, \tilde{\mu}_{ik})$, for $j, k = 1, \ldots, n_i$ and this quantity is calculated as follows.

$$\begin{aligned} Cov(y_{ij}, \tilde{\mu}_{ik}) &= Cov(y_{ij}, exp(x'_{ik}(t_{ik})\beta + \hat{\gamma}(t_{ik})) \\ &= Cov\left(y_{ij}, exp(x'_{ik}(t_{ik})\beta)\ \frac{\sum_{l=1}^{K}\sum_{u=1}^{n_l} w_{lu}(t_{ik})y_{lu}}{\sum_{l=1}^{K}\sum_{u=1}^{n_l} w_{lu}(t_{ik})exp(x'_{lu}\beta)}\right) \\ &= \frac{exp(x'_{ik}(t_{ik})\beta)}{\sum_{l=1}^{K}\sum_{u=1}^{n_l} w_{lu}(t_{ik})exp(x'_{lu}\beta)} Cov(y_{ij}, \sum_{l=1}^{K}\sum_{u=1}^{n_l} w_{lu}(t_{ik})y_{lu}) \end{aligned}$$

Under the assumption that y_{ij}'s are independent, $Cov(y_{ij}, \sum_{l=1}^{K}\sum_{u=1}^{n_l} w_{lu}(t_{ik})y_{lu}) = Cov(y_{ij}, w_{ij}(t_{ik})y_{ij})$ and this implies

$$Cov(y_{ij}, \tilde{\mu}_{ik}) = \frac{exp(x'_{ik}(t_{ik})\beta)}{\sum_{l=1}^{K}\sum_{u=1}^{n_l} w_{lu}(t_{ik})exp(x'_{lu}\beta)} \sum_{u=1}^{n_i} w_{iu}(t_{ik})\sigma^{(ns)}_{iju}.$$

Note that by using the formula for $\Sigma_i^{*(ns)}(\rho)$ from (3.27) and the derivative formula from (3.26), we solve the FSSGQL(ns) estimating equation (3.25) using the Newton-Raphson method. Letting

$$f(\beta) = \sum_{i=1}^{K} \frac{\partial E(\tilde{\mu}_i')}{\partial \beta} \; [\Sigma_i^{*(ns)}]^{-1} \; (y_i - \tilde{\mu}_i),$$

and starting with an initial value of β, the iterative equation

$$\hat{\beta}_{(r+1)} = \hat{\beta}_{(r)} - [(f'(\beta))^{-1} f(\beta)]_{\beta=\beta_{(r)}} \tag{3.28}$$

updates the value of β in each step until convergence.

3.2.2.5 Existing PSSGEE approach

Instead of using the estimating equation given in (3.18), authors, such as Severini and Staniswalis (1994) and Lin and Carroll (2001) use different estimating equation to estimate β, which has the form

$$\sum_{i=1}^{K} \frac{\partial \tilde{\mu}_i'}{\partial \beta} \; [V_i(\alpha)]^{-1} \; (y_i - \tilde{\mu}_i) \;\; = \;\; 0, \tag{3.29}$$

where $V_i(\alpha)$ is computed as

$$\begin{aligned} V_i(\hat{\alpha}) = \hat{var}(Y_i) \;\; &= \;\; A_i^{1/2} \; R_i(\hat{\alpha}) \; A_i^{1/2} \\ &= \;\; A_i^{1/2} \; R(\hat{\alpha}) \; A_i^{1/2} \text{ for the case } n_i = n \end{aligned} \tag{3.30}$$

with $R(\hat{\alpha})$ as the constant stationary 'working' correlation matrix. The estimating equation in (3.29) is referred to as the GEE, but because it uses $var(Y_i)$ instead of $var(Y_i - \tilde{\mu}_i)$, for clarity we refer to this equation as the partially standardized semi-parametric GEE (PSSGEE).

There are two problems when using $R_i(\alpha)$ in estimating $var(Y_i)$. First, in the non-stationary case, the correlations should be dependent on the time dependent covariates. Thus, using a stationary version, say $C_i(\rho)$ (3.3), for the true non-stationary correlation matrix $C_i^{(ns)}(\rho)$ is an approximation. Secondly, $R_i(\alpha)$ is not only stationary but its form also may differ from $C_i(\rho)$ as $R_i(\alpha)$ is a user's choice matrix. In addition, there is no guidance for choosing $R_i(\alpha)$ and in the longitudinal setup with fully specified regression function, it was shown by Sutradhar and Das (1999) [see also Sutradhar (2010, 2011)] that use of $R_i(\alpha)$ may produce inconsistent [Crowder (1995)] or consistent but inefficient estimates for β as compared to the simpler moment or QL approaches. As a remedy to this problem, Sutradhar (2003) proposed a GQL approach which always produces efficient estimates compared to the independence correlation based GEE approach. Thus, it seems appropriate to examine the effects of GEE estimates for β obtained from (3.29) by comparing with the PSSGQL(ns) approach under the present semi-parametric setup. These comparison studies are performed through various simulations and the results are provided in the next Chapter.

3.2.2.6 Estimation of 'working' correlation parameter α

Following the existing GEE methods, we use the estimating equation in (3.29) with $\hat{var}(Y_i) = A_i^{1/2}\ R(\hat{\alpha})\ A_i^{1/2}$, for the estimation of the regression parameter β. Similar to the linear model case, the 'working' correlation matrix $R(\alpha)$ is estimated under various correlation structures, namely, AR(1), MA(1), EQC, independence (I) and unstructured (UNS) assumptions. The 'working' correlation parameter α, for these correlation structures is estimated by solving the respective moment equations. For

example, for EQC correlations structure,

$$\hat{\alpha} = \frac{\sum_{i=1}^{K} \sum_{j \neq u}^{n_i} y_{ij}^* y_{iu}^*}{\sum_{i=1}^{K} \sum_{j=1}^{n_i} (y_{ij}^*)^2} \tag{3.31}$$

where $y_{ij}^* = \frac{y_{ij} - exp(x_{ij}'(t_{ij})\hat{\beta} + \hat{\gamma}(t_{ij}))}{\sqrt{exp(x_{ij}'(t_{ij})\hat{\beta} + \hat{\gamma}(t_{ij}))}}$ and for AR(1) and MA(1) correlation structures, $\hat{\alpha}$ is computed by using

$$\hat{\alpha} = \frac{\sum_{i=1}^{K} \sum_{j=1}^{n_i - 1} y_{ij}^* y_{i,j+1}^*}{\sum_{i=1}^{K} \sum_{j=1}^{n_i} (y_{ij}^*)^2}. \tag{3.32}$$

Under the unstructured correlation structure [Lin and Carroll (2001)], $R_i(\alpha)$ is estimated as

$$\hat{R}(\alpha) = \frac{1}{K} \sum_{i=1}^{K} r_i r_i',$$

where $r_i = (r_{i1}, \ldots, r_{ij}, \ldots, r_{in_i})'$ is the vector of residuals with $r_{ij} = \frac{y_{ij} - exp(x_{ij}'(t_{ij})\hat{\beta} + \hat{\gamma}(t_{ij}))}{\sqrt{exp(x_{ij}'(t_{ij})\hat{\beta} + \hat{\gamma}(t_{ij}))}}$.

3.3 Semi-parametric longitudinal models for binary data with non-stationary correlation structures

Let $y_{ij}(t_{ij})$ be the j^{th} binary response for i^{th} individual at time point t_{ij}. In the binary case, the marginal properties of the model are different than that for count data model. A typical choice for the marginal mean would be a logit function. Thus, we write the mean and variance of the binary model as

$$E(Y_{ij}|x_{ij}) = \mu_{ij} = \frac{exp[x_{ij}'(t_{ij})\beta + \gamma(t_{ij})]}{1 + exp[x_{ij}'(t_{ij})\beta + \gamma(t_{ij})]} \tag{3.33}$$

$$V(Y_{ij}|x_{ij}) = \sigma_{ijj} = \frac{exp[x_{ij}'(t_{ij})\beta + \gamma(t_{ij})]}{(1 + exp[x_{ij}'(t_{ij})\beta + \gamma(t_{ij})])^2} \tag{3.34}$$

for $i = 1, \ldots, K; j = 1, \ldots, n_i$, where K is the number of individuals, and n_i is the number of responses for individual i.

3.3.1 Non-stationary correlation models for binary data

The construction of the non-stationary correlation matrix $C_i^{(ns)}(\rho)$ for the semi-parametric longitudinal count data was given in Section 3.1.2 under three different such as AR(1), MA(1) and EQC structures. Note that the formulas for the lag correlations under the binary models would be different than the count data models. Following Sutradhar (2010), the next three subsections provide the non-stationary correlation structures under the binary AR(1), MA(1) and EQC models.

3.3.1.1 Non-stationary AR(1) models in semi-parametric setup

The non-stationary AR(1) type model for the binary responses y_{ij} under the semi-parametric setup has the probability relationship

$$\begin{aligned} P[Y_{i1}=1|x_{i1}] &= \mu_{i1} \\ P[Y_{i1}=1|y_{i,j-1},x_{ij},x_{i,j-1}] &= \mu_{ij}+\rho(y_{i,j-1}-\mu_{i,j-1}) \text{ for } j=2,\ldots,n_i, \end{aligned}$$

where the mean μ_{ij} in terms of the non-parametric function is given by

$$\mu_{ij}=\frac{exp[x_{ij}'(t_{ij})\beta+\gamma(t_{ij})]}{1+exp[x_{ij}'(t_{ij})\beta+\gamma(t_{ij})]}.$$

It then follows that

$$V(Y_{ij}|x_{ij}) = \mu_{ij}(1-\mu_{ij})=\frac{exp[x_{ij}'(t_{ij})\beta+\gamma(t_{ij})]}{(1+exp[x_{ij}'(t_{ij})\beta+\gamma(t_{ij})])^2}.$$

The covariance between the responses y_{ij} and y_{ik} can be derived in a similar way using the conditioning and unconditioning principle that we have used under the count data model. To be specific,

$$cov(Y_{ij},Y_{ik}|x_{ij},x_{ik}) = E(Y_{ij}Y_{ik}|x_{ij},x_{ik})-\mu_{ij}\mu_{ik}$$

Hence for $j < k$,

$$cov(Y_{ij}, Y_{ik}|x_{ij}, x_{ik}) = \rho^{k-j}\mu_{ij}(1-\mu_{ij}),$$

and it then follows that correlation between the responses y_{ij} and y_{ik} has the formula

$$\begin{aligned} corr(Y_{ij}, Y_{ik}|x_{ij}, x_{ik}) &= c^{(ns)}_{i,j,k}(x_{ij}, x_{ik}.\rho) \\ &= \begin{cases} \rho^{k-j}\sqrt{\frac{\mu_{ij}(1-\mu_{ij})}{\mu_{ik}(1-\mu_{ik})}} & j < k \\ \rho^{j-k}\sqrt{\frac{\mu_{ik}(1-\mu_{ik})}{\mu_{ij}(1-\mu_{ij})}} & j > k \end{cases} \end{aligned} \tag{3.35}$$

with the range restriction

$$max\left[-\frac{\mu_{ij}}{1-\mu_{i,j-1}}, -\frac{1-\mu_{ij}}{\mu_{i,j-1}},\right] \leq \rho \leq min\left[\frac{1-\mu_{ij}}{1-\mu_{i,j-1}}, \frac{\mu_{ij}}{\mu_{i,j-1}},\right].$$

But, when the model follows a stationary correlation structure, the correlations in (3.35) reduce to $\rho^{|k-j|}$.

3.3.2 Non-stationary MA(1) models in semi-parametric setup

Under the non-stationary MA(1) correlation structure, the binary responses follow a probability relationship

$$\begin{aligned} P[Y_{i1} = 1|x_{i1}] &= \mu_{i1} \\ P[Y_{i1} = 1|d_{ij}, d_{i,j-1}] &= d_{ij} + \rho d_{i,j-1} \text{ for } j = 2, \ldots, n_i, \end{aligned}$$

where d_{ij}'s are independently distributed with the following mean and variance [Sutradhar (2010)]

$$\begin{aligned} E(d_{ij}) &= \sum_{u=0}^{j-1}(-\rho)^u\mu_{i,j-u} \\ V(d_{ij}) &= \left(\frac{\sum_{u=0}^{j-1}(-\rho)^u\mu_{i,j-u}}{\sum_{u=0}^{j-1}(-\rho)^u}\right)\left(1 - \frac{\sum_{u=0}^{j-1}(-\rho)^u\mu_{i,j-u}}{\sum_{u=0}^{j-1}(-\rho)^u}\right) \end{aligned}$$

where

$$\mu_{ij} = \frac{exp[x'_{ij}(t_{ij})\beta + \gamma(t_{ij})]}{1 + exp[x'_{ij}(t_{ij})\beta + \gamma(t_{ij})]} \quad \text{and} \quad V(Y_{ij}|x_{ij}) = \mu_{ij}(1 - \mu_{ij}).$$

Next, for $j < k$, the correlation between the responses y_{ij} and y_{ik} is given by

$$\begin{aligned} corr(Y_{ij}, Y_{ik}|x_{ij}, x_{ik}) &= c^{(ns)}_{i,j,k}(x_{ij}, x_{ik}, \rho) \qquad (3.36) \\ &= \begin{cases} \dfrac{\rho\left(\frac{\sum_{u=0}^{j-1}(-\rho)^u \mu_{i,j-u}}{\sum_{u=0}^{j-1}(-\rho)^u}\right)\left(1 - \frac{\sum_{u=0}^{j-1}(-\rho)^u \mu_{i,j-u}}{\sum_{u=0}^{j-1}(-\rho)^u}\right)}{\sqrt{\mu_{ij}(1-\mu_{ij})}\,\sqrt{\mu_{ik}(1-\mu_{ik})}} & \text{for } k-j=1 \\ 0 & \text{for } k-j>1 \end{cases} \end{aligned}$$

However, under the stationary model, the correlations have the simple formula given by

$$corr(Y_{ij}, Y_{ik}) = \begin{cases} \rho & \text{for } |k-j| = 1 \\ 0 & \text{otherwise} \end{cases}$$

3.3.3 Non-stationary EQC models in semi-parametric setup

Under the non-stationary EQC model, the probability relationship for the responses may be written as

$$P[Y_{ij} = 1|y_{i0}, x_{ij}] = \mu_{ij} + \rho(y_{i0} - \mu_{i1}) \ , i = 1, \ldots, K; \ j = 1, \ldots, n_i \qquad (3.37)$$

with

$$\mu_{ij} = \frac{exp[x'_{ij}(t_{ij})\beta + \gamma(t_{ij})]}{1 + exp[x'_{ij}(t_{ij})\beta + \gamma(t_{ij})]}$$

Also, in (3.37) it is assumed that y_{i0} is an unobservable initial binary response, which has the same mean as y_{i1}. Thus, we can write the mean and variance of this model

as

$$\begin{aligned} E(Y_{ij}|x_{ij}) &= E_{y_{i0}} E(Y_{ij}|y_{i0}) \\ &= E(\mu_{ij} + \rho(y_{i0} - \mu_{i1})) = \mu_{ij} \\ Var(Y_{ij}|x_{ij}) &= E_{y_{i0}}[Var(Y_{ij}|y_{i0})] + Var_{y_{i0}}[E(Y_{ij}|y_{i0})] \\ &= \mu_{ij}(1-\mu_{ij}). \end{aligned}$$

Next, the covariance and correlations are derived for $j \neq k$ as follows.

$$\begin{aligned} cov(Y_{ij}, Y_{ik}|x_{ij}, x_{ik}) &= E_{y_{i0}}[cov(Y_{ij}, Y_{ik}|y_{i0})] \\ &\quad + cov_{y_{i0}}[(\mu_{ij} + \rho(y_{i0} - \mu_{i1}), (\mu_{ik} + \rho(y_{i0} - \mu_{i1})] \\ &= \rho^2 \mu_{i0}(1-\mu_{i0}) = \rho^2 \mu_{i1}(1-\mu_{i1}) \qquad (3.38) \\ corr(Y_{ij}, Y_{ik}|x_{ij}, x_{ik}) &= c^{(ns)}_{i,j,k}(x_{ij}, x_{ik}, \rho) \\ &= \frac{\rho^2 \mu_{i1}(1-\mu_{i1})}{\sqrt{\mu_{ij}(1-\mu_{ij})}\sqrt{\mu_{ik}(1-\mu_{ik})}} \qquad (3.39) \end{aligned}$$

Note that if the binary data follow a stationary correlation structure, the non-stationary correlations in (3.39) reduce to

$$corr(Y_{ij}, Y_{ik}) = \rho^2.$$

3.4 Estimation in semi-parametric models in longitudinal binary data

3.4.1 Estimation of non-parametric function $\gamma(.)$

To estimate the non-parametric function, we use the SQL approach discussed in Section 1.2.3 for the binary data in the independence setup. For a given value of

β, say, $\hat{\beta}$ the SQL estimating equation for $\gamma(t_0)$ at a particular time point t_0 can be written as

$$\sum_{i=1}^{K}\sum_{j=1}^{n_i} w_{ij}(t_0)\ \frac{\partial \mu_{ij}}{\partial \gamma(t_0)}\left[\frac{y_{ij}-\mu_{ij}}{\mu_{ij}(1-\mu_{ij})}\right] = 0$$

where

$$\mu_{ij} = \frac{exp[x'_{ij}(t_{ij})\hat{\beta} + \gamma(t_0)]}{1 + exp[x'_{ij}(t_{ij})\hat{\beta} + \gamma(t_0)]}.$$

Because

$$\frac{\partial \mu_{ij}}{\partial \gamma(t_0)} = \mu_{ij}(1-\mu_{ij}),$$

the above estimating equation reduces to

$$\sum_{i=1}^{K}\sum_{j=1}^{n_i} w_{ij}(t_0)\ (y_{ij}-\mu_{ij}) = 0 \tag{3.40}$$

where $w_{ij}(t_0) = \frac{p_{ij}(\frac{t_0-t_{ij}}{b})}{\sum_{i=1}^{K}\sum_{j=1}^{n_i} p_{ij}(\frac{t_0-t_{ij}}{b})}$, $p_{ij}(\frac{t_0-t_{ij}}{b}) = \frac{1}{\sqrt{2\pi}b} exp(\frac{-1}{2}(\frac{t_0-t_{ij}}{b})^2)$, b is the bandwidth parameter and t_{ij} is the time measure for the i^{th} individual at time point j.

Unlike the count or linear model cases, the estimating equation (3.40) does not provide a closed form formula for $\gamma(t_0)$. Thus, we use the Newton-Raphson method to solve (3.40). For a known value of β, say $\hat{\beta}$, we denote the estimating function in the left-hand side of (3.40) as

$$f(\gamma(t_0), \hat{\beta}) = \sum_{i=1}^{K}\sum_{j=1}^{n_i} w_{ij}(t_0)\ (y_{ij}-\mu_{ij}), \tag{3.41}$$

and write the Newton-Raphson iterative equation as

$$\hat{\gamma}_{(k+1)} = \hat{\gamma}_{(k)} - \left([f'_{\gamma}(\gamma, \hat{\beta})]^{-1} f(\gamma, \hat{\beta})\right)_{\gamma=\hat{\gamma}_{(k)}} \tag{3.42}$$

to obtain, for example, the improved value at $(k+1)^{th}$ iteration using the value from the k^{th} iteration. The iteration then continues until convergence. The derivative

function $f'_\gamma(\gamma(t_0), \hat{\beta})$ in (3.42) has the formula

$$f'_\gamma(\gamma(t_0), \hat{\beta}) \;=\; -\sum_{i=1}^{K}\sum_{j=1}^{n_i} w_{ij}(t_0)\mu_{ij}(1-\mu_{ij}). \tag{3.43}$$

3.4.1.1 PSSGQL(ns) estimation of β

For the estimation of regression parameter β, by considering

$$\tilde{\mu}_{ij} = \frac{exp(x_{ij}(t_{ij})\beta + \hat{\gamma}(t_{ij}, \beta))}{[1 + exp(x_{ij}(t_{ij})\beta + \hat{\gamma}(t_{ij}, \beta))]}, \tag{3.44}$$

similar to that of (3.18) for count data, we can write the partially standardized semi-parametric GQL(ns) (PSSGQL(ns)) estimating equation under the longitudinal binary model setup as

$$\sum_{i=1}^{K} \frac{\partial \tilde{\mu}'_i}{\partial \beta}\; [var(Y_i)]^{-1}\; (y_i - \tilde{\mu}_i) \;=\; 0, \tag{3.45}$$

where $\tilde{\mu}_i = [\tilde{\mu}_{i1}, \ldots, \tilde{\mu}_{ij}, \ldots \tilde{\mu}_{in_i}]'$ with $\tilde{\mu}_{ij}$ defined as in (3.44) and the variance function has the form

$$\hat{var}(Y_i) = \Sigma_i^{(ns)}(\hat{\rho}) = A_i^{1/2}\; C_i^{(ns)}(\hat{\rho})\; A_i^{1/2}, \tag{3.46}$$

where $C_i^{(ns)}(\hat{\rho})$ can be computed for a known correlation model discussed in Section 3.3.1. The elements in the gradient functions are calculated as follows.

$$\begin{aligned} \frac{\partial \tilde{\mu}_{ij}}{\partial \beta} &= \frac{exp(x_{ij}(t_{ij})\beta + \hat{\gamma}(t_{ij}, \beta))}{[1 + exp(x_{ij}(t_{ij})\beta + \hat{\gamma}(t_{ij}, \beta))]^2}\; [x_{ij}(t_{ij}) + \frac{\partial \hat{\gamma}(t_{ij}, \beta)}{\partial \beta}] \\ &= \tilde{\mu}_{ij}(1 - \tilde{\mu}_{ij})[x_{ij}(t_{ij}) + \frac{\partial \hat{\gamma}(t_{ij}, \beta)}{\partial \beta}]. \end{aligned} \tag{3.47}$$

The quantity $\frac{\partial\hat{\gamma}(t_{ij},\beta)}{\partial\beta}$ in (3.47) can be calculated by using the concept of Taylor's series expansion and it then follows from (3.42) that

$$\begin{aligned}
\frac{\partial\hat{\gamma}(t_{ij},\beta)}{\partial\beta} &= -\frac{\partial}{\partial\beta}\left([f'_\gamma(\gamma(t_{ij}),\beta)]^{-1}f(\gamma(t_{ij}),\beta)\right) \\
&= -\left([f'_\gamma(\gamma(t_{ij}),\beta)]^{-1}\frac{\partial}{\partial\beta}f(\gamma(t_{ij}),\beta)+\frac{\partial}{\partial\beta}\,[f'_\gamma(\gamma(t_{ij}),\beta)]^{-1}\,[f(\gamma(t_{ij}),\beta)]\right) \\
&= -\,[f'_\gamma(\gamma(t_{ij}),\beta)]^{-1}\left(\frac{\partial}{\partial\beta}f(\gamma(t_{ij}),\beta)\right) \\
&\quad +[f'_\gamma(\gamma(t_{ij}),\beta)]^{-1}\,\left(\frac{\partial}{\partial\beta}\,[f'_\gamma(\gamma(t_{ij}),\beta)]\right)\,[f'_\gamma(\gamma(t_{ij}),\beta)]^{-1}\,[f(\gamma(t_{ij}),\beta)].
\end{aligned}$$

where by (3.41),

$$\begin{aligned}
\frac{\partial}{\partial\beta}f(\gamma(t_{ij}),\beta) &= \frac{\partial}{\partial\beta}\sum_{l=1}^{K}\sum_{u=1}^{n_l}w_{lu}(t_{ij})\;(y_{lu}-\tilde{\mu}_{lu}) \\
&= -\sum_{l=1}^{K}\sum_{u=1}^{n_l}w_{lu}(t_{ij})\tilde{\mu}_{lu}(1-\tilde{\mu}_{lu})[x_{lu}(t_{lu})+\frac{\partial\hat{\gamma}(t_{lu})}{\partial\beta}]
\end{aligned}$$

and by (3.43),

$$\begin{aligned}
\frac{\partial}{\partial\beta}f'_\gamma(\gamma(t_{ij}),\beta) &= \frac{\partial}{\partial\beta}\left(-\sum_{l=1}^{K}\sum_{u=1}^{n_l}w_{lu}(t_{ij})\tilde{\mu}_{lu}(1-\tilde{\mu}_{lu})\right) \\
&= -\sum_{l=1}^{K}\sum_{u=1}^{n_l}w_{lu}(t_{ij})\tilde{\mu}_{lu}(1-\tilde{\mu}_{lu})[x_{lu}(t_{lu})+\frac{\partial\hat{\gamma}(t_{lu})}{\partial\beta}](1-2\tilde{\mu}_{lu}).
\end{aligned}$$

Analogous to the estimation procedure explained in the longitudinal count data model, there is no closed-form expression for β in the current binary setup as well. In fact, in the present binary case it is more complicated to obtain a form for $\frac{\partial\hat{\gamma}(t_{ij},\beta)}{\partial\beta}$. Nevertheless, one can use Newton-Raphson method to solve the estimating equation (3.45). The iterative equation for the this method is given by

$$\hat{\beta}_{(k+1)}=\hat{\beta}_{(k)}-\left([f'_\beta(\hat{\gamma}(t,\beta),\beta)]^{-1}f(\hat{\gamma}(t,\beta),\beta)\right)_{\beta=\hat{\beta}_{(k)}} \tag{3.48}$$

where

$$f(\hat{\gamma}(t,\beta),\beta)=\sum_{i=1}^{K}\frac{\partial\tilde{\mu}'_i}{\partial\beta}\;\;[\Sigma_i^{(ns)}(\hat{\rho})]^{-1}\;(y_i-\tilde{\mu}_i)$$

and

$$f'_{\beta}(\hat{\gamma}(t,\beta),\beta) \;=\; -\sum_{i=1}^{K} \frac{\partial \tilde{\mu}'_i}{\partial \beta}\; [\Sigma_i^{(ns)}(\hat{\rho})]^{-1}\; \frac{\partial \tilde{\mu}_i}{\partial \beta'}.$$

3.4.1.2 Estimation of correlation index parameter ρ

Similar to the calculations under the count data model, we equate the sample covariance with its population counterpart as

$$\frac{\sum_{i=1}^{K}\sum_{j=2}^{n_i} y^*_{ij}y^*_{i,j-1}}{\sum_{i=1}^{K}(n_i-1)} \;=\; \frac{1}{\sum_{i=1}^{K}(n_i-1)} \sum_{i=1}^{K}\sum_{j=2}^{n_i} \rho\; \frac{\sqrt{\sigma_{i,j-1}}}{\sqrt{\sigma_{ij}}}.$$

This yields the moment estimating equation for ρ under the non-stationary AR(1) correlation model as

$$\hat{\rho} = \frac{\sum_{i=1}^{K}\sum_{j=2}^{n_i} y^*_{ij}y^*_{i,j-1}}{\sum_{i=1}^{K}\sum_{j=1}^{n_i} y^{*2}_{ij}}\; \frac{\sum_{i=1}^{K} n_i}{\sum_{i=1}^{K}\sum_{j=2}^{n_i}\left[\frac{\hat{\sigma}_{i,j\;\;1}}{\hat{\sigma}_{ij}}\right]^{\frac{1}{2}}}, \tag{3.49}$$

where $y^*_{ij} = \frac{y_{ij}-\hat{\mu}_{ij}}{\sqrt{\hat{\sigma}_{ij}}}$ with $\hat{\mu}_{ij} = \frac{exp(x'_{ij}(t_{ij})\hat{\beta}+\hat{\gamma}(t_{ij}))}{[1+exp(x'_{ij}(t_{ij})\hat{\beta}+\hat{\gamma}(t_{ij}))]}$ and $\hat{\sigma}_{ij} = \hat{\mu}_{ij}(1-\hat{\mu}_{ij})$. And, the estimates $\hat{\beta}$ and $\hat{\gamma}(t_{ij})$ are computed by using the PSSGQL(ns) and SQL approaches for β and $\gamma(t_{ij})$, respectively.

3.4.1.3 FSSGQL(ns) estimation of β

Similar to the FSSGQL(ns) estimating equation (3.25) for β in the count data case, the estimating equation for the binary case is given by

$$\sum_{i=1}^{K} \frac{\partial E(\tilde{\mu}'_i)}{\partial \beta}\; [\Sigma_i^{*(ns)}]^{-1}\; (y_i - \tilde{\mu}_i) = 0, \tag{3.50}$$

where the elements $\tilde{\mu}_{ij}$ in $\tilde{\mu}_i$ has the form given in (3.44) whereas in the count data case $\tilde{\mu}_{ij} = exp(x_{ij}(t_{ij})\beta + \hat{\gamma}(t_{ij},\beta))$. Note that because of the difference in formulas for $\tilde{\mu}_i$, $\hat{\gamma}(.)$ in $\tilde{\mu}_{ij}$ can not be obtained explicitly for the binary data. This makes the

computation for $\frac{\partial E(\tilde{\mu}_i')}{\partial \beta}$ as well as $\Sigma_i^{*(ns)}$ extremely complicated. However, the empirical study to be discussed in Chapter 4 demonstrated that for the count data case, the FSSGQL(ns) offers a slight improvement only over the estimation by the PSSGQL(ns) approach. It is expected that this pattern might be true for the binary case as well. For this reason, we have not pursued the exact computations for the components of the estimating equation (3.50). Further investigations for any approximation may be useful but is not attempted at this stage.

Chapter 4

Empirical Study for Semi-parametric Longitudinal Count Data Models

We have provided a detailed finite sample based numerical study in Chapter 2 under the Gaussian-type ARMA auto-correlation models involving a semi-parametric regression function. It was found that the proposed FSSGQL approach produces uniformly more efficient regression estimates than the existing PSSGEE approaches. In this chapter we examine the finite sample performances of the aforementioned approaches for the discrete data case. More specifically, we choose the count data models for the empirical study because of the fact that the semi-parametric numerical analysis for such longitudinal count data is not adequately discussed in the literature.

The organization of the empirical study in this chapter is as follows. In Sections 4.1 and 4.2, we provide the simulation design and data generation. Section 4.3 ex-

amines the performance of the naive GQL (NGQL) approach which shows the extent of bias in the estimation of β when the non-parametric function is ignored in the estimation. In Section 4.4, we compare the relative efficiency performance of the proposed PSSGQL(ns) approach with the existing PSSGEE approaches. We also study the performance of the FSSGQL(ns) approach in Section 4.5.

4.1 Simulation design

(a) Sample Size: $K = 100$; $n_i = 4$ for $i = 1, \ldots, K$; and $t_{ij} = j$ for all $i = 1, \ldots, K$, and $j = 1, \ldots, n_i$.

(b) Covariate Selection: Similar to the fully specified longitudinal model studied in Sutradhar (2010), we consider $p = 2$ time dependent covariates with their values as

$$x_{ij1}(t_{ij}) = \begin{cases} \frac{1}{2} & \text{for } i = 1, \ldots, 25 \text{ and } j = 1, 2 \\ 1 & \text{for } i = 1, \ldots, 25 \text{ and } j = 3, 4 \\ \frac{-1}{2} & \text{for } i = 26, \ldots, 75 \text{ and } j = 1 \\ 0 & \text{for } i = 26, \ldots, 75 \text{ and } j = 2, 3 \\ \frac{1}{2} & \text{for } i = 26, \ldots, 75 \text{ and } j = 4 \\ \frac{j}{2n_i} & \text{for } i = 76, \ldots, 100 \text{ and } j = 1, 2, 3, 4 \end{cases}$$

$$x_{ij2}(t_{ij}) = \begin{cases} \frac{j-2.5}{2n_i} & \text{for } i = 1, \ldots, 50 \text{ and } j = 1, 2, 3, 4 \\ 0 & \text{for } i = 51, \ldots, 100 \text{ and } j = 1, 2 \\ \frac{1}{2} & \text{for } i = 51, \ldots, 100 \text{ and } j = 3, 4 \end{cases}$$

Note that the covariate values are chosen to reflect the variable time dependence

for the different groups of individuals. Thus, the choice is quite general. One may choose other specific covariates depending on the situations.

(c) Covariate Effects: We choose three different sets of covariate effects.

(1) $(\beta_1, \beta_2)' = (0, 0)'$

(2) $(\beta_1, \beta_2)' = (0.5, 0.5)'$

(3) $(\beta_1, \beta_2)' = (1.0, 1.0)'$

(d) Nonparametric function: We consider a quadratic function for $\gamma(t_{ij})$ as

$$\gamma(t_{ij}) = 0.3 + 0.2\ (t_{ij} - \frac{n_i + 1}{2}) + 0.05\ (t_{ij} - \frac{n_i + 1}{2})^2; \quad n_i = 4$$

which is similar as that of the linear model case considered in the simulation study in Section 2.3. Note that this function is unknown in practice. Hence for the inferences this is treated as a non-parametric function.

4.2 Data generation

We choose the semi-parametric AR(1) non-stationary correlation model to generate the data. To be specific, for all $i = 1, \ldots, 100$ and $j = 1, \ldots 4$, y_{ij}'s are generated as follows.

(a) y_{i1} is generated using $y_{i1} \sim Poi(\mu_{i1} = exp(x'_{i1}(t_{i1})\beta + \gamma(t_{i1})))$ where $x'_{i1}(t_{i1})$ and $\gamma(.)$ are given under the simulation design.

(b) For $t_{ij} = j = 2, \ldots, 4$, y_{ij}'s are generated following the binomial thinning operation $\rho * y_{i,j-1} = \sum_{s=1}^{y_{i,j}\ 1} b_s(\rho)$ with $d_{ij} \sim Poi(\mu_{ij} - \rho\mu_{i,j-1})$.

4.3 NGQL estimation: A biased approach

To obtain NGQL estimate of β, we solve the NGQL estimating equation (3.17) which was constructed by ignoring $\gamma(.)$ in the mean response function. The data are generated following Section 4.2 and the simulations are repeated for 1000 times. The computational steps for NSGQL estimation is as follows.

Step 1. Starting with an initial value of β and an initial value of correlation index parameter ρ, we solve (3.17) to obtain the NGQL estimate of β.

Step 2. We estimate ρ from (3.22) using the estimate of β from **Step 1**.

Step 3. Repeat **Steps 1** and **2** in order to obtain improved estimates for β and ρ.

The simulation results are provided in the Table 4.1. As expected, the estimates of β are biased for various choice of the regression parameter β and ρ. For example, for the true regression parameter $\beta = (0.5, 0.5)'$, the estimated value of β when $\rho = 0.8$ is $(1.0318, 1.2595)'$, which shows very large bias in estimating β by using $\hat{\beta}_{NGQL}$.

Table 4.1: Simulated means (SMs), simulated standard errors (SSEs) and mean squared error (MSEs) of the naive estimates of regression parameters β under non-stationary AR(1) correlation model for selected values of correlation index parameter ρ with K=100; n=4; and 1000 simulations.

True $\beta = (\beta_1, \beta_2)'$	ρ	Qunantity	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\rho}$
$\beta = (0, 0)'$	0.2	SM	0.4595	0.8190	0.2011
		SSE	0.1063	0.1820	0.0624
		MSE	0.2224	0.7039	
	0.5	SM	0.4747	0.7620	0.4532
		SSE	0.1158	0.1796	0.0620
		MSE	0.2387	0.6129	
	0.8	SM	0.4894	0.6903	0.7014
		SSE	0.1018	0.1684	0.0448
		MSE	0.2499	0.5086	
$\beta = (0.5, 0.5)'$	0.2	SM	1.0000	1.3233	0.1793
		SSE	0.0826	0.1442	0.0622
		MSE	0.2568	0.6986	
	0.5	SM	1.0072	1.2979	0.4090
		SSE	0.0875	0.1486	0.0595
		MSE	0.2649	0.6587	
	0.8	SM	1.0318	1.2595	0.6378
		SSE	0.0884	0.1432	0.0482
		MSE	0.2906	0.5973	

Table 4.1 Continued

True β	ρ	Qunantity	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\rho}$
$\beta = (1,1)'$	0.2	SM	1.5076	1.8333	0.1667
		SSE	0.0634	0.1188	0.0609
		MSE	0.2617	0.7083	
	0.5	SM	1.5199	1.8217	0.3712
		SSE	0.0685	0.1181	0.0597
		MSE	0.2750	0.6891	
	0.8	SM	1.5312	1.8150	0.5795
		SSE	0.0708	0.1221	0.0534
		MSE	0.2872	0.6791	

4.4 A finite sample efficiency comparison between PSSGQL(ns) and PSSGEE estimations

Because the NGQL estimates were found to be highly biased, we now proceed to examine the performance of the proposed PSSGQL(ns) and other existing PSSGEE approaches, which are developed by considering that the regression function contains a non-parametric function as well. For the simulation studies, we generate correlated count data as described in Section 4.2 along with three different values of correlation index parameter. To consider both low and high correlations, we have chosen $\rho = 0.2, 0.5$ and 0.8. The simulations are repeated for 1000 times. For the estimation of β including ρ and the non-parametric function $\gamma(.)$, we follow the following steps.

Step 1. For an initial value of β, we solve the 'working' independence assumption based SQL estimating equation (3.15) to estimate the non-parametric function $\gamma(\cdot)$.

Step 2. Starting with an initial value of β, $\hat{\gamma}(\cdot)$ from **Step 1**, and an initial value of correlation index parameter ρ, we use (3.21) to obtain the proposed PSSGQL(ns) estimate of β.

Step 3. Next, we estimate ρ from (3.22) using the estimates of $\gamma(\cdot)$ and β from **Steps 1** and **2**, respectively.

Step 4. We repeat **Steps 1**, **2** and **3** in order to obtain improved estimates for the non-parametric function $\gamma(\cdot)$, β and ρ.

The computational steps for PSSGEE approaches are the same as above, except that in these approaches, the 'working' correlation parameter α is computed depending on the chosen correlation structure. For convenience, we denote PSSGEE(AR(1)), PSSGEE(MA(1)), PSSGEE(EQC), PSSGEE(I), PSSGEE(UNS) to

represent the PSSGEE approaches under 'working' correlation structures AR(1), MA(1), EQC, independence and unstructured respectively. We consider the mean squared error (MSE) for this comparison study. The simulation results for three different sets of regression parameters namely, $(\beta_1, \beta_2)' = [(0,0)', (1,1)', (0.5, 0.5)']$ are provided in the Tables 4.2, 4.3 and 4.4 respectively, for the proposed PSSGQL(ns) and PSSGQL(s), as well as for the existing PSGEE approaches.

The results from Table 4.2 show that for a selected set of true values of $\beta = (\beta_1, \beta_2)' = (0,0)'$, the MSE under the proposed PSSGQL(ns) are uniformly smaller than the PSSGEE approaches for $\rho = 0.2, 0.5$ or 0.8. This pattern also holds when $(\beta_1, \beta_2)' = (1,1)'$ as displayed in Table 4.3. However, when $\beta = (\beta_1, \beta_2)' = (0.5, 0.5)'$, some of the PSSGEE methods appear to work as good as PSSGQL(ns) for low correlation case. Turning back to Table 4.2, when PSSGQL(ns) regression estimates are compared to that of PSSGQL(s), the MSEs under PSSGQL(ns) are uniformly smaller than those under PSSGQL(s), as expected. The difference between the MSEs is significant when correlations are large. However, when PSSGQL(s) and PSSGEE approaches are compared, PSSGQL(s) appear to perform almost the same as the PSSGEE(AR(1)), PSSGEE(MA(1)) and PSSGEE(UNS), but PSSGEE(EQC) and PSSGEE(I) perform the worst. To illustrate these relative performances, we point out, for example, the MSEs of all approaches when correlation is large. More specifically it follows from Table 4.2 with $\beta = (0,0)'$ and $\rho = 0.8$, the MSE for β_2 estimate under PSSGQL(ns) is 0.0891 followed by 0.1620 for PSSGEE(UNS), and the MSE for the worse case PSSGEE(I) being 0.2029. It appears from these results that there can be a huge efficiency loss in the main regression parameter estimation when PSSGEE(I) or other PSSGEE methods are used, especially when data are highly correlated.

When estimating β, we have to estimate the non-parametric function $\gamma(.)$ involved in the semi-parametric regression function (3.1). As discussed in Section 3.2.1, $\gamma(.)$ is estimated by using the SQL approach for known β. Since β is estimated by using various PSSGQL and PSSGEE methods, $\gamma(.)$ is also estimated under each of these methods. The resulting estimates of $\hat{\gamma}(.)$ under these different methods along with the true $\gamma(.)$ function are displayed in Figures 4.1, 4.2 and 4.3 for large correlation cases. In the estimation procedure for estimating non-parametric function, we have used the bandwidth $b = \frac{1}{(4K)^{1/5}}$ [Pagan and Ullah (1999)], for example. It can be seen from the figures that the non-parametric function is estimated well for different regression parameter values. This is because all the estimated functions appear to be close to the true curve for the selected non-parametric function. The results are similar for the other cases.

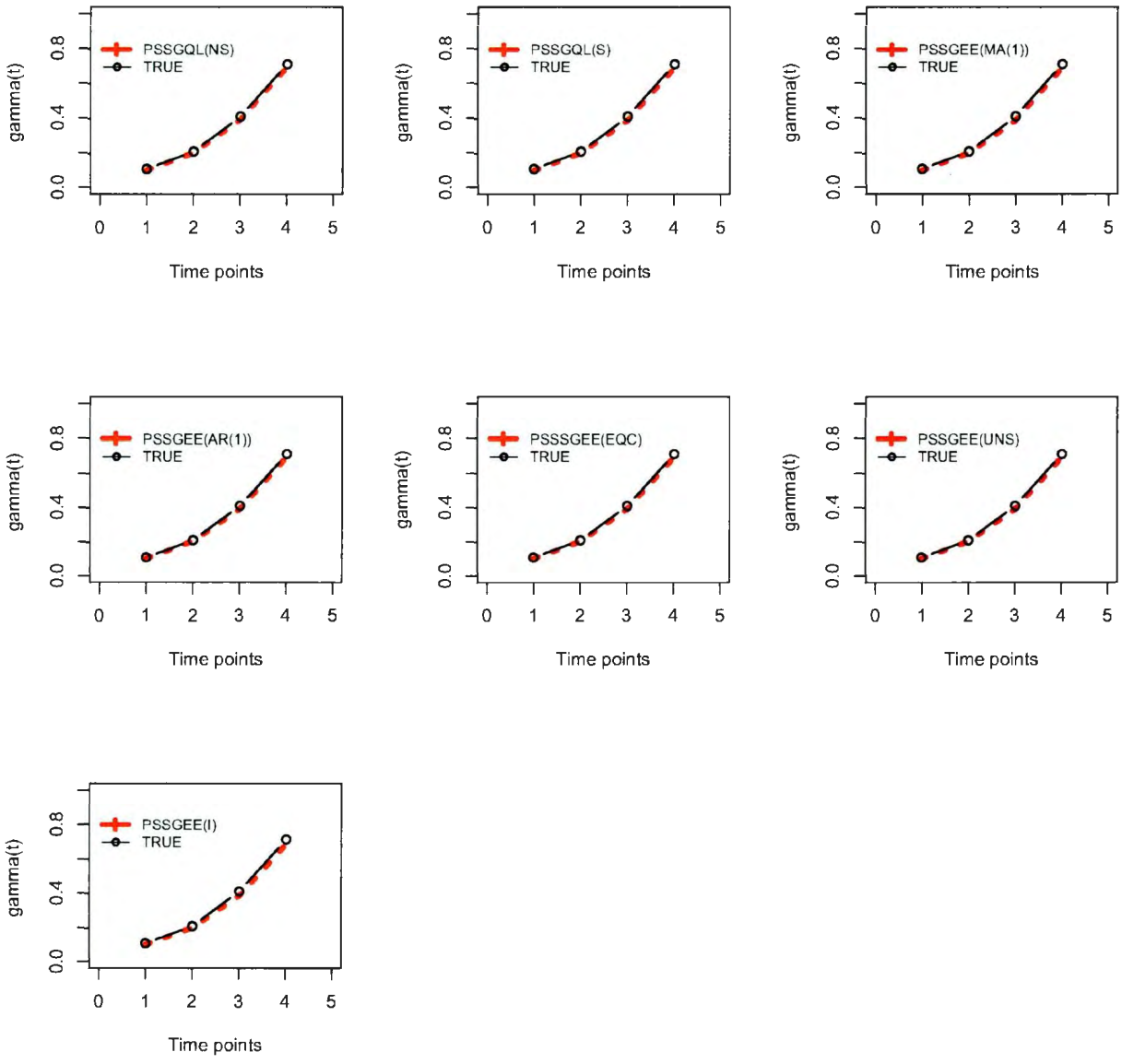

Figure 4.1: Simulated means of estimates of $\gamma(t)$ for PSSGQL and PSSGEE methods, and true values of $\gamma(t)$ under non-stationary AR(1) correlation models for count data with a correlation index parameter $\rho = 0.8$ and regression parameters $(\beta_1, \beta_2)' = (0, 0)'$.

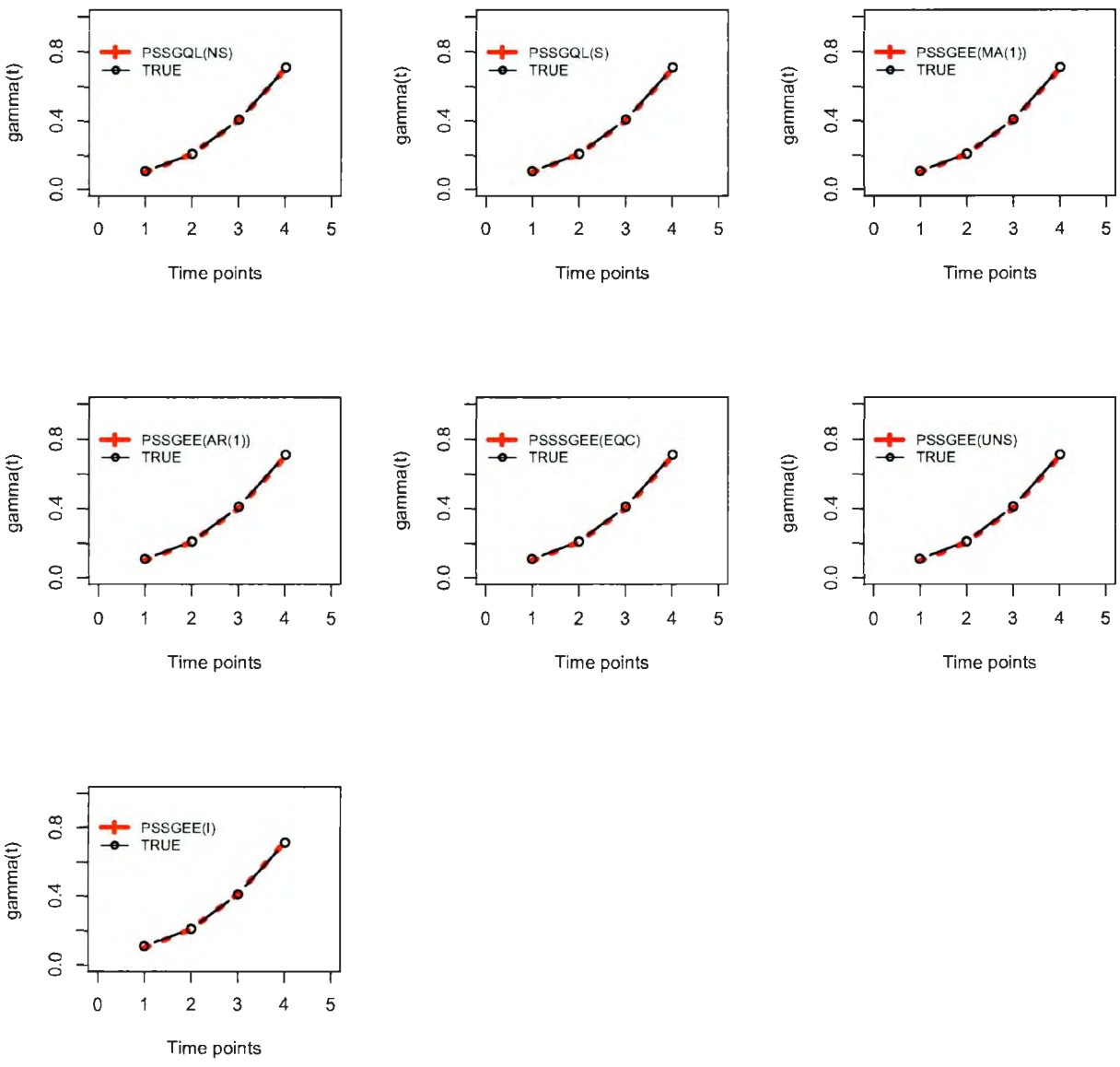

Figure 4.2: Simulated means of estimates of $\gamma(t)$ for PSSGQL and PSSGEE methods, and true values of $\gamma(t)$ under non-stationary AR(1) correlation models for count data with a correlation index parameter $\rho = 0.8$ and regression parameters $(\beta_1, \beta_2)' = (1, 1)'$.

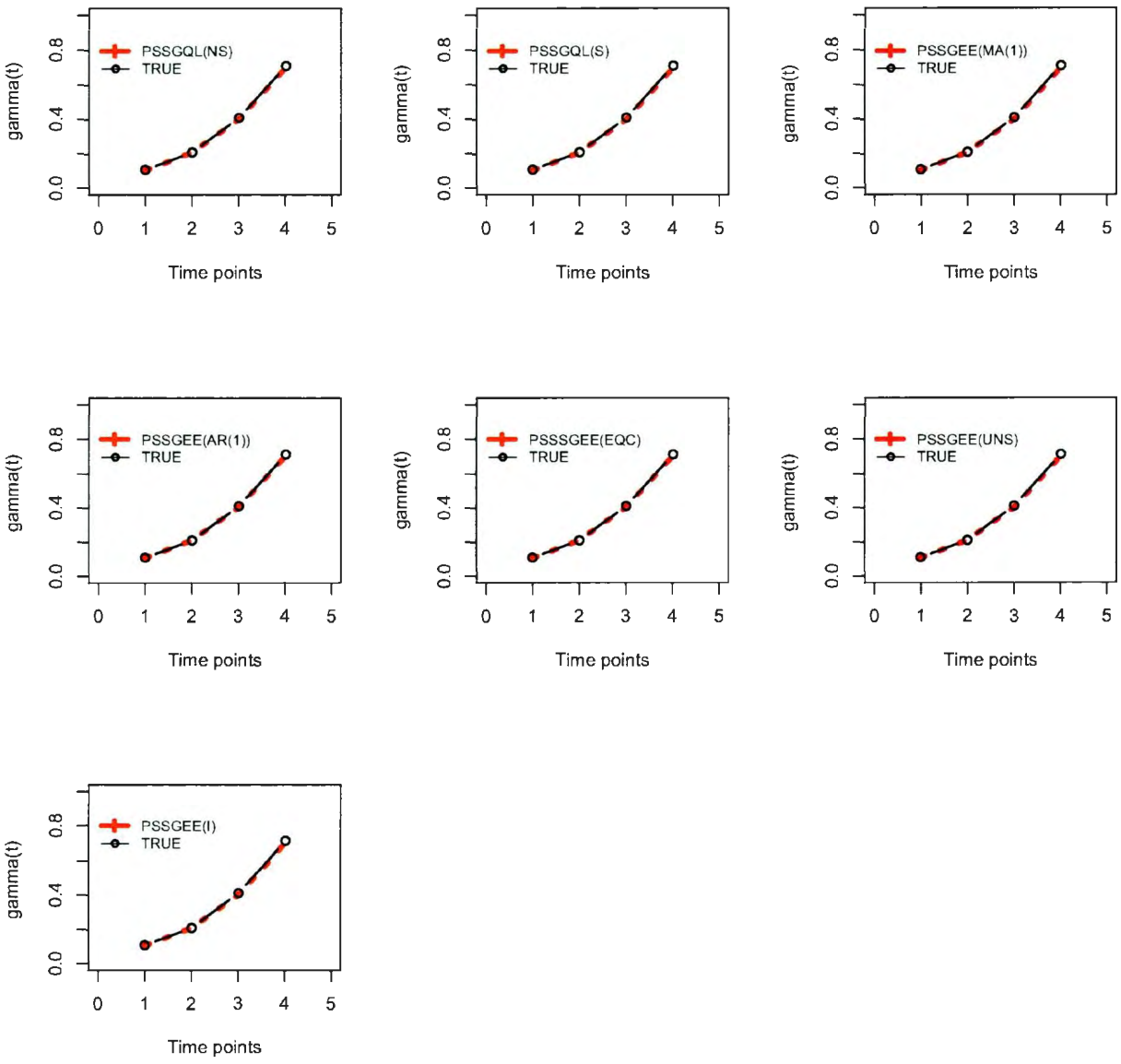

Figure 4.3: Simulated means of estimates of $\gamma(t)$ for PSSGQL and PSSGEE methods, and true values of $\gamma(t)$ under non-stationary AR(1) correlation models for count data with a correlation index parameter $\rho = 0.8$ and regression parameters $(\beta_1, \beta_2)' = (0.5, 0.5)'$.

Table 4.2: Simulated means (SMs), simulated standard errors (SSEs) and mean squared error (MSEs) of the PSSGQL and PSSGEE estimates of regression parameters $\beta_1 = 0.0$ and $\beta_2 = 0.0$, under non-stationary AR(1) correlation model for selected values of correlation index parameter ρ with K=100; n=4; and 1000 simulations.

ρ	Method	Quantity	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\alpha}$	$\hat{\rho}$	$\hat{\rho}_1$	$\hat{\rho}_2$	$\hat{\rho}_3$
0.2	PSSGQL(ns)	SM	-0.0158	-0.0102		0.1776			
		SSE	0.1533	0.3221		0.0597			
		MSE	0.0238	0.1039					
	PSSGQL(s)	SM	-0.0162	-0.0108			0.1769	0.0256	-0.0014
		SSE	0.1537	0.3264			0.0594	0.0756	0.1023
		MSE	0.0239	0.1067					
	PSSGEE (AR(1))	SM	-0.0161	-0.0105	0.1769				
		SSE	0.1534	0.3262	0.0594				
		MSE	0.0238	0.1065					
	PSSGEE (MA(1))	SM	-0.0163	-0.0106	0.1769				
		SSE	0.1533	0.3262	0.0594				
		MSE	0.0238	0.1065					
	PSSGEE (EQC)	SM	-0.0163	-0.0109					
		SSE	0.1546	0.3269					
		MSE	0.0242	0.1070					
	PSSGEE (I)	SM	-0.0187	-0.0148					
		SSE	0.1576	0.3285					
		MSE	0.0252	0.1081					
	PSSGEE (UNS)	SM	-0.0170	-0.0108					
		SSE	0.1536	0.3265					
		MSE	0.0239	0.1067					

Table 4.2 Continued

ρ	Method	Quantity	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\alpha}$	$\hat{\rho}$	$\hat{\rho}_1$	$\hat{\rho}_2$	$\hat{\rho}_3$
0.5	PSSGQL(ns)	SM	0.0062	0.0234		0.4494			
		SSE	0.1671	0.3228		0.0582			
		MSE	0.0280	0.1047					
	PSSGQL(s)	SM	0.0060	0.0215			0.4473	0.1970	0.0837
		SSE	0.1753	0.3488			0.0580	0.0781	0.0989
		MSE	0.0308	0.1221					
	PSSGEE (AR(1))	SM	0.0059	0.0215	0.4473				
		SSE	0.1751	0.3491	0.0580				
		MSE	0.0307	0.1223					
	PSSGEE (MA(1))	SM	0.0058	0.0213	0.4472				
		SSE	0.1740	0.3497	0.0579				
		MSE	0.0303	0.1223					
	PSSGEE (EQC)	SM	0.0065	0.0218	0.3033				
		SSE	0.1823	0.3543	0.0604				
		MSE	0.0333	0.1260					
	PSSGEE (I)	SM	0.0018	0.0158					
		SSE	0.1919	0.3622					
		MSE	0.0368	0.1314					
	PSSGEE (UNS)	SM	0.0057	0.0214					
		SSE	0.1754	0.3478					
		MSE	0.0308	0.1214					
0.8	PSSGQL(ns)	SM	0.0071	0.0125		0.7177			
		SSE	0.1549	0.2982		0.0431			
		MSE	0.0240	0.0891					
	PSSGQL(s)	SM	0.0106	0.0157			0.7139	0.5076	0.3617
		SSE	0.1858	0.4079			0.0430	0.0687	0.0940
		MSE	0.0346	0.1666					
	PSSGEE (AR(1))	SM	0.0107	0.0155	0.7140				
		SSE	0.1852	0.4076	0.0430				
		MSE	0.0344	0.1664					
	PSSGEE (MA(1))	SM	0.0109	0.0153	0.7137				
		SSE	0.1820	0.4095	0.0430				
		MSE	0.0332	0.1679					
	PSSGEE (EQC)	SM	0.0107	0.0159	0.5864				
		SSE	0.1982	0.4181	0.0543				
		MSE	0.0394	0.1751					
	PSSGEE (I)	SM	0.0164	0.0167					
		SSE	0.2261	0.4501					
		MSE	0.0512	0.2029					
	PSSGEE (UNS)	SM	0.0103	0.0158					
		SSE	0.1859	0.4022					
		MSE	0.0347	0.1620					

Table 4.3: Simulated means (SMs), simulated standard errors (SSEs) and mean squared error (MSEs) of the PSSGQL and PSSGEE estimates of regression parameters $\beta_1 = 1.0$ and $\beta_2 = 1.0$, under non-stationary AR(1) correlation model for selected values of correlation index parameter ρ with K=100; n=4; and 1000 simulations.

ρ	Method	Quantity	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\alpha}$	$\hat{\rho}$	$\hat{\rho}_1$	$\hat{\rho}_2$	$\hat{\rho}_3$
0.2	PSSGQL (ns)	SM	1.0033	1.0048		0.1470			
		SSE	0.1300	0.2409		0.0598			
		MSE	0.0169	0.0581					
	PSSGQL(s)	SM	1.0033	1.0052			0.1465	0.0152	-0.0015
		SSE	0.1307	0.2424			0.0596	0.0730	0.1020
		MSE	0.0171	0.0588					
	PSSGEE (AR(1))	SM	1.0032	1.0049	0.1465				
		SSE	0.1306	0.2427	0.0596				
		MSE	0.0171	0.0589					
	PSSGEE (MA(1))	SM	1.0032	1.0050	0.1465				
		SSE	0.1305	0.2427	0.0596				
		MSE	0.0170	0.0589					
	PSSGEE (EQC)	SM	1.0031	1.0051	0.0780				
		SSE	0.1314	0.2431	0.0474				
		MSE	0.0173	0.0591					
	PSSGEE (I)	SM	0.9982	0.9978					
		SSE	0.1507	0.2643					
		MSE	0.0227	0.0699					
	PSSGEE (UNS)	SM	0.9726	0.9839					
		SSE	0.3571	0.5081					
		MSE	0.1283	0.2584					

Table 4.3 Continued

ρ	Method	Quantity	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\alpha}$	$\hat{\rho}$	$\hat{\rho}_1$	$\hat{\rho}_2$	$\hat{\rho}_3$
0.5	PSSGQL (ns)	SM	1.0020	1.0032		0.3690			
		SSE	0.1400	0.2541		0.0576			
		MSE	0.0196	0.0646					
	PSSGQL(s)	SM	1.0013	1.0004			0.3679	0.1277	0.0422
		SSE	0.1447	0.2708			0.0573	0.0747	0.1020
		MSE	0.0209	0.0733					
	PSSGEE (AR(1))	SM	1.0009	1.0000	0.3679				
		SSE	0.1450	0.2711	0.0573				
		MSE	0.0210	0.0735					
	PSSGEE (MA(1))	SM	1.0010	1.0001	0.3679				
		SSE	0.1445	0.2717	0.0573				
		MSE	0.0209	0.0738					
	PSSGEE (EQC)	SM	1.0012	1.0007	0.2334				
		SSE	0.1475	0.2715	0.0558				
		MSE	0.0218	0.0737					
	PSSGEE (I)	SM	1.0064	1.0010					
		SSE	0.1658	0.2861					
		MSE	0.0275	0.0819					
	PSSGEE (UNS)	SM	0.9837	0.9820					
		SSE	0.4717	0.6122					
		MSE	0.2228	0.3751					
0.8	PSSGQL(ns)	SM	0.9997	1.0160		0.5966			
		SSE	0.1479	0.2658		0.0501			
		MSE	0.0219	0.0709					
	PSSGQL(s)	SM	1.0022	1.0188			0.5945	0.3409	0.2115
		SSE	0.1668	0.3083			0.0499	0.0766	0.0996
		MSE	0.0278	0.0954					
	PSSGEE (AR(1))	SM	1.0024	1.0192	0.5945				
		SSE	0.1672	0.3084	0.0499				
		MSE	0.0280	0.0955					
	PSSGEE (MA(1))	SM	1.0023	1.0189	0.5944				
		SSE	0.1640	0.3091	0.0499				
		MSE	0.0269	0.0959					
	PSSGEE (EQC)	SM	1.0023	1.0192	0.4460				
		SSE	0.1746	0.3124	0.0590				
		MSE	0.0305	0.0980					
	PSSGEE (I)	SM	0.9874	0.9982					
		SSE	0.1852	0.3127					
		MSE	0.0345	0.0978					
	PSSGEE (UNS)	SM	0.9742	1.0179					
		SSE	0.5872	0.5472					
		MSE	0.3455	0.2997					

Table 4.4: Simulated means (SMs), simulated standard errors (SSEs) and mean squared error (MSEs) of the PSSGQL and PSSGEE estimates of regression parameters $\beta_1 = 0.5$ and $\beta_2 = 0.5$, under non-stationary AR(1) correlation model for selected values of correlation index parameter ρ with K=100; n=4; and 1000 simulations.

ρ	Method	Quantity	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\alpha}$	$\hat{\rho}$	$\hat{\rho}_1$	$\hat{\rho}_2$	$\hat{\rho}_3$
0.2	PSSGQL(ns)	SM	0.5110	0.5109		0.1586			
		SSE	0.1381	0.2755		0.0601			
		MSE	0.0192	0.0760					
	PSSGQL(s)	SM	0.5106	0.5120			0.1581	0.0234	0.0024
		SSE	0.1381	0.2756			0.0599	0.0743	0.1013
		MSE	0.0192	0.0761					
	PSSGEE (AR(1))	SM	0.5106	0.5120	0.1581				
		SSE	0.1380	0.2757	0.0599				
		MSE	0.0192	0.0762					
	PSSGEE (MA(1))	SM	0.5108	0.5121	0.1581				
		SSE	0.1380	0.2757	0.0599				
		MSE	0.0192	0.0762					
	PSSGEE (EQC)	SM	0.5105	0.5119	0.0872				
		SSE	0.1385	0.2755	0.0494				
		MSE	0.0193	0.0760					
	PSSGEE (I)	SM	0.5153	0.5119					
		SSE	0.1511	0.2859					
		MSE	0.0231	0.0819					
	PSSGEE (UNS)	SM	0.5109	0.5115					
		SSE	0.1386	0.2764					
		MSE	0.0193	0.0765					

Table 4.4 Continued

ρ	Method	Quantity	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\alpha}$	$\hat{\rho}$	$\hat{\rho}_1$	$\hat{\rho}_2$	$\hat{\rho}_3$
0.5	PSSGQL(ns)	SM	0.4950	0.5028		0.40588			
		SSE	0.1551	0.2969		0.0565			
		MSE	0.0241	0.0882					
	PSSGQL(s)	SM	0.4963	0.5056			0.4042	0.1566	0.0628
		SSE	0.1607	0.3208			0.0561	0.0774	0.1006
		MSE	0.0258	0.1029					
	PSSGEE (AR(1))	SM	0.4958	0.5051	0.4042				
		SSE	0.1607	0.3207	0.0561				
		MSE	0.0258	0.1029					
	PSSGEE (MA(1))	SM	0.4960	0.5053	0.4041				
		SSE	0.1593	0.3211	0.0561				
		MSE	0.0254	0.1031					
	PSSGEE (EQC)	SM	0.4958	0.5056	0.2647				
		SSE	0.1654	0.3228	0.0568				
		MSE	0.0274	0.1042					
	PSSGEE (I)	SM	0.4959	0.5048					
		SSE	0.1748	0.3218					
		MSE	0.0306	0.1036					
	PSSGEE (UNS)	SM	0.4961	0.5055					
		SSE	0.1611	0.3194					
		MSE	0.0260	0.1020					
0.8	PSSGQL(ns)	SM	0.5076	0.5079		0.6529			
		SSE	0.1485	0.2753		0.0462			
		MSE	0.0221	0.0759					
	PSSGQL(s)	SM	0.5098	0.5085			0.6503	0.4182	0.2776
		SSE	0.1717	0.3456			0.0461	0.0714	0.0947
		MSE	0.0296	0.1195					
	PSSGEE (AR(1))	SM	0.5095	0.5083	0.6503				
		SSE	0.1715	0.3456	0.0461				
		MSE	0.0295	0.1195					
	PSSGEE (MA(1))	SM	0.5103	0.5084	0.6502				
		SSE	0.1671	0.3464	0.0460				
		MSE	0.0280	0.1201					
	PSSGEE (EQC)	SM	0.5093	0.5092	0.5108				
		SSE	0.1827	0.3523	0.0555				
		MSE	0.0335	0.1242					
	PSSGEE (I)	SM	0.5109	0.5156					
		SSE	0.1904	0.3602					
		MSE	0.0364	0.1300					
	PSSGEE (UNS)	SM	0.5117	0.5119					
		SSE	0.1792	0.3538					
		MSE	0.0322	0.1253					

4.5 Performance of the FSSGQL(ns) estimation

Recall from Chapter 3 that in addition to the PSSGQL estimation, we also proposed the FSSGQL approach (Section 3.2.2.4) when estimation effect of $\gamma(.)$ is accommodated in the longitudinal weight matrix to construct the estimating equation. In this section, we examine whether the FSSGQL(ns) approach offers any improvement over the PSSGQL approach for longitudinal count data. The data generation and estimation steps are similar to that in Section 4.4. More specifically, the estimation steps are:

Step 1. For an initial value of β, we solve the 'working' independence assumption based SQL estimating equation (3.15) to estimate the non-parametric function $\gamma(\cdot)$.

Step 2. Starting with an initial value of β, $\hat{\gamma}(\cdot)$ from **Step 1**, and an initial value of correlation index parameter ρ, we use (3.28) to obtain the proposed FSSGQL(ns) estimate of β.

Step 3. Next, we estimate ρ from (3.22) using the estimates of $\gamma(\cdot)$ and β from **Steps 1** and **2**, respectively.

Step 4. We repeat **Steps 1**, **2** and **3** in order to obtain improved estimates for the non-parametric function $\gamma(\cdot)$, β and ρ.

The simulation results for the FSSGQL(ns) approach are given in Table 4.5. We have also displayed the non-parametric function estimates ($\hat{\gamma}(.)$) using $\hat{\beta}_{FSSGQL(ns)}$ in Figures 4.4 and 4.5 for $\beta = (0, 0)'$ and $\beta = (0.5, 0.5)'$ respectively. The figures show that the non-parametric function is estimated well. As far as the estimation of the main regression parameter β is concerned, FSSGQL(ns) appears to perform almost the same, offering in general slight reduction in the MSEs as compared to

the PSSGQL(ns) approach. For example, when $\beta = (0,0)'$ and $\rho = 0.5$, the MSEs for $\hat{\beta}_1$ and $\hat{\beta}_2$ under the FSSGQL approach are 0.0222 and 0.0783 whereas under the PSSGQL(ns) approach they are 0.0280 and 0.1047 respectively.

Table 4.5: Simulated means (SMs), simulated standard errors (SSEs) and mean squared error (MSEs) of the FSSGQL(ns) estimates of regression parameter β under non-stationary AR(1) correlation model for selected values of correlation index parameter ρ with K=100; n=4; and 1000 simulations.

True $\beta = (\beta_1, \beta_2)'$	ρ	Quantity	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\rho}$
$\beta = (0,0)'$	0.5	SM	0.0049	0.0178	0.4477
		SSE	0.1489	0.2792	0.0579
		MSE	0.0222	0.0783	
	0.8	SM	0.0102	0.0107	0.7145
		SSE	0.1572	0.3258	0.0430
		MSE	0.0248	0.1063	
$\beta = (0.5,0.5)'$	0.5	SM	0.4457	0.3986	0.4049
		SSE	0.1399	0.2739	0.0560
		MSE	0.0225	0.0853	
	0.8	SM	0.4534	0.3899	0.6510
		SSE	0.1484	0.2939	0.0460
		MSE	0.0242	0.0985	

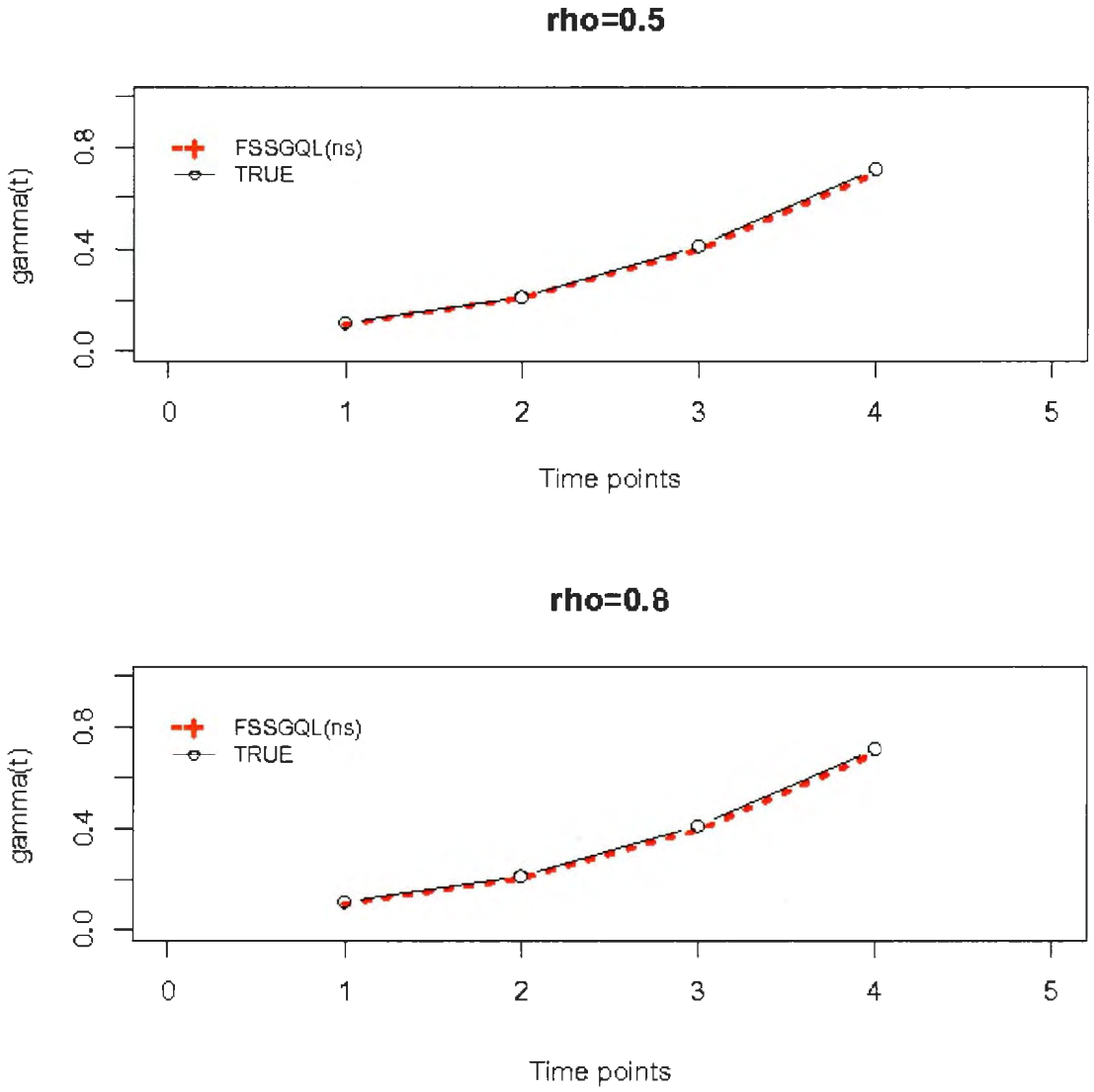

Figure 4.4: Simulated means of estimates of $\gamma(t)$ for FSSGQL(ns) method and true values of $\gamma(t)$ under non-stationary AR(1) correlation models for count data with regression parameters $(\beta_1, \beta_2)' = (0, 0)'$.

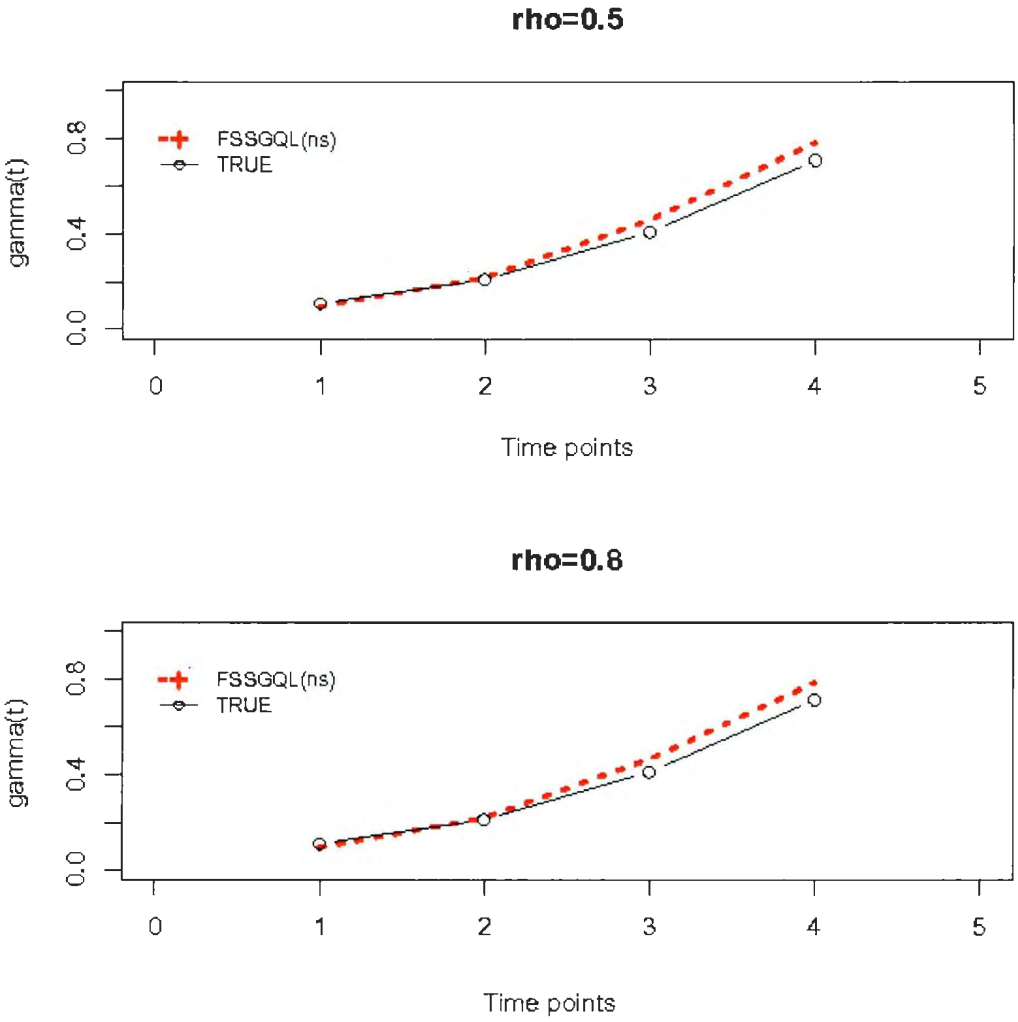

Figure 4.5: Simulated means of estimates of $\gamma(t)$ for FSSGQL(ns) method and true values of $\gamma(t)$ under non-stationary AR(1) correlation models for count data with regression parameters $(\beta_1, \beta_2)' = (0.5, 0.5)'$.

Chapter 5

Concluding Remarks

When the regression function is fully specified, there exists GEE and GQL approaches in the longitudinal setup for efficient estimation of the regression parameters. As opposed to the GEE approach, the GQL approach is developed for a class of Gaussian-type auto-correlation models. It is known that the GEE approach may produce less efficient regression estimates as compared to the independence assumption based QL or MM approaches, whereas the GQL approach produces more efficient estimates. In this thesis, we have studied the semi-parametric regression models where the regression function also contains a non-parametric function in the longitudinal setup for both continuous and discrete data. It is found that similar to the completely longitudinal setup, the SGQL (semi-parametric GQL) approach produces uniformly more efficient regression estimates than the SGEE (semi-parametric GEE) approaches, including the independence assumption based SGEE(I) approach. This is demonstrated in the linear model setup in Chapter 2, and for longitudinal count data in Chapter 4.

Unlike some of the existing SGEE approaches, in this thesis we have estimated

the non-parametric function based on the independence assumption, whereas the regression effects are estimated by exploiting the non-stationary correlation structure of the repeated discrete responses. Furthermore, as opposed to the existing SGEE approaches, we have accommodated the estimation effect of the non-parametric function while estimating the regression parameters. This resulted in the FSSGQL (fully standardized SGQL) and PSSGQL (partially standardized SGQL) approaches. The performances of all these approaches are discussed in details in Chapter 2 for continuous correlated data, and in Chapters 3 and 4 for discrete correlated data. We found that in the linear model setup, the FSSGQL approach yielded uniformly more efficient regression estimates than the PSSGEE approaches. In the discrete data setup, the PSSGQL approach produced more efficient estimates than the PSSGEE approaches. Also, the FSSGQL approach provided slightly more efficient regression estimates than the PSSGQL approach.

While this thesis has provided useful inferences for generalized linear longitudinal semi-parametric models, future research should investigate an approximation to ease the computation aspects in the semi-parametric longitudinal binary data setup. Further research should investigate the modelling of correlations when responses are collected based on unequi-spaced time points. Also, in the longitudinal setup, it may happen that a portion of the data is missing at random. The semi-parametric inference for such missing data would be of interest to researchers, presenting more complicated inferences.

Bibliography

[1] Ahmad, I. A. and Lin, P. E. (1976). Non-parametric sequential estimation of a multiple regression function. *Bulletin of Mathematical Statistics*, **17**, 63-75.

[2] Altman, N.S, (1990). Kernel Smoothing of Data with Correlated Errors. *Journal of the American Statistical Association*, **85**, 749-758.

[3] Amemiya, T. (1985). *Advanced Econometrics.* Cambridge, MA: Harvard University Press.

[4] Andrews, D.W.K. (1995). Nonparametric Kernel Estimation for Semiparametric Models. *Econometric Theory*, **11**, 560-596.

[5] Ansley, C.F., Kohn, R. and Tharm, D. (1991). The performance of cross-validation and maximum likelihood estimators of spline smoothing parameters, *Journal of the American Statistical Association*, **86**, 1042-50.

[6] Bierens, H. J. (1987). *Kernel estimators of regression functions*, Advances in Econometrics, Cambridge University Press.

[7] Carota, C. and Parmigiani, G. (2002). Semiparametric regression for count data. *Biometrika*, **89**, 265281.

[8] Casella, George, and Berger, Roger L. (1990). *Statistical Inference*, Duxbury Press, U.S.A.

[9] Cleveland, W.S. (1979). Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association*, **74**, 829-836.

[10] Craven, P. and G. Wahba (1979). Smoothing noisy data with spline functions. *Numerische Mathematik*, **31**, 377-403.

[11] Crowder, M. (1995). On the use of a working correlation matrix in using generalized linear models for repeated measures. *Biometrika*, **82**, 407-10.

[12] Eubank, R. (1988). *Spline smoothing and nonparametric regression*, Dekker, New York.

[13] Fan, J. (1992). Design-adaptive nonparametric regression. *Journal of the American Statistical Association*,**87**, 998-1004.

[14] Fan, J. (1993). Local linear smoothers and their minimax efficiency. *The Annals of Statistics*, **21**, 196-216.

[15] Fan, J., Huang, T. and Li, R. (2007). Analysis of longitudinal data with semiparametric estimation of covariance function. *Journal of the American Statistical Association*, **102**, 632-641.

[16] Fan, J and Wu, Y. (2008). Semi-parametric estimation of covariance matrices for longitudinal data. *Journal of the American Statistical Association*, **103**, 1520-1533.

[17] Fitzmaurice, G. M., Laird, N. M. and Rotnitzky, A. G. (1993). Regression models for discrete longitudinal responses (with discussion). *Statistical Science*, **8**, 284-309.

[18] Gilmour, A. R., Anderson, R. D., and Rae, A. L. (1985). The analysis of binomial data by a generalized linear mixed model. *Biometrika*, **72**, 593-599.

[19] Greblicki, W. and Krzyzak, A. (1980). Asymptotic properties of kernel estimates of a regression function. *Journal of Statistical Planning and Inference*, **4**, 81-90.

[20] Hastie, T. and Tibshirani, R. (1990). Generalized Additive Models. *London: Chapman and Hall.*

Journal of the Royal Statistical Society (B), **55**, 757796.

[21] Li, Y. (2011). Efficient semiparametric regression for longitudinal data with nonparametric covariance estimation. *Biometrika*, **98**, 355-370.

[22] Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13-22.

[23] Lin, X. and Carroll, R. J. (2001). Semi-parametric regression for clustered data. *Biometrika*, **88**, 1179-1185.

[24] Lin, X. and Carroll, R. J. (2001a). Semi-parametric regression for clustered data using generalized estimating equations. *Journal of the American Statistical Association*, **96**, 1045-1056.

[25] McCullagh, P. (1983). Quasi-Likelihood Functions. *The Annals of Statistics*, **11**, 59-67.

[26] McCullagh, P., and Nelder, J. A. (1989). Generalized linear models (2nd ed.). *Chapman and Hall*, London.

[27] Moyeed, R.A. and Diggle, P.J. (1994). Rates of convergence in semi-parametric modelling of longitudinal data. *The Australian Journal of Statistics*,**36**, 75-93.

[28] Muller, H.G. (1988). Nonparametric Regression Analysis of Longitudinal Data. *Lecture Notes in Statistics*, **46**, Springer-Verlag, New York.

[29] Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability and its Applications*, **9** 141-142.

[30] Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A*, **135**, 370384.

[31] Pagan, A. and Ullah, A. (1999). *Nonparametric Econometrics, Cambridge University Press*, Cambridge.

[32] Picard, R. and Cook, R. D. (1984). Cross-validation of regression models. *Journal of the American Statistical Association*, **79**, 575583.

[33] Rao, C. R. (1973). *Linear Statistical Inference and Its Applications.* New York: John Wiley & Sons.

[34] Royall, R. M. (1966). A class of nonparametric estimates of a smooth regression function, doctoral dissertation, Stanford University.

[35] Severini, T. A., and Staniswalis, J. G. (1994). Quasi-likelihood estimation in semi-parametric models. *Journal of the American Statistical Association*, **89**, 501-511.

[36] Severini, T. A., and Wong, W. H. (1992). Generalized profile likelihood and conditionally parametric models. *The Annals of Statistics*, **20**, 1768-1802.

[37] Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London.

[38] Stone, C. J. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society*, **36**, 111-147.

[39] Stone, C. J. (1977). Consistent nonparametric regression. *The Annals of Statistics*, **5**, 595-645.

[40] Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *The Annals of Statistics*, **8**, 1348-1360.

[41] Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, **10**, 1348-1360.

[42] Sneddon, G. and Sutradhar, B.C. (2004). On semi-parametric familial-longitudinal models. *Statistics and Probability Letters*, **69**, 369-379.

[43] Speckman, P. E. (1988). Regression analysis of partially linear models. *Journal of the Royal Statistical Society*, Ser. B, **50**, 413-436.

[44] Staniswalis, J. G. (1989). On the kernel estimate of a regression function in likelihood-based models. *Journal of the American Statistical Association*, **84**, 276-283.

[45] Sutradhar, B.C. and Das, K. (1999). On the efficiency of regression estimators in generalized linear models for longitudinal data. *Biometrika*, **86**, 459-465.

[46] Sutradhar, B. C. (2003). An overview on regression models for discrete longitudinal responses. *Statistical Science*, **18**, 377-393.

[47] Sutradhar, B.C. (2010). Inferences in generalized linear longitudinal mixed models. *Canadian Journal of Statistics*, **38**, 174-196.

[48] Sutradhar, B.C. (2010a). Generalized Quasi-likelihood (GQL) Inference (version 8). StatProb: The Encyclopedia Sponsored by Statistics and Probability Societies. Freely available at *http://statprob.com/encyclopedia/GeneralizedQuasiLikelihoodGQLInferences.html.*

[49] Sutradhar, B. C. (2011). *Dynamic Mixed Models for Familial Longitudinal Data.* New York: Springer.

[50] Wahba, G. (1990).*Spline Models for Observational Data.* SIAM, Philadelphia.

[51] Watson, G.S. (1964). Smooth regression analysis. *Sankhya Series A*, **26**, 359-372.

[52] Wang, N., Carroll, R. J. and Lin, X. (2005). Efficient semi-parametric marginal estimation for longitudinal/clustered data. *Journal of the American Statistical Association*, **100**, 147-157.

[53] Wedderburn, R.W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*, **61**, 439-447.

[54] Whittaker, E.T. (1923). On a new method of graduation. *Proceedings of the Edinburgh Mathematical Society*, **41**, 62-75.

[55] You, J. and Chen, G. (2007). Semi-parametric generalized least squares estimation in partially linear regression models with correlated errors, *Journal of Statistical Planning and Inference*, **137**, 117-132.

[56] ZegerS. L., Liang,K. Y, and Albert. P. S (1988). Models for longitudinal data: A generalized estimating equation approach, *Biometrics*, **44**, 1049-1060.

[57] Zeger, S. L. and Liang, K. Y., (1986). Longitudinal data analysis for discrete and continuous outcomes, *Biometrics*, **42**, 121-130.

[58] Zeger, S. L., and Diggle, P. J. (1994). Semi-parametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters, *Biometrics*, **50**,689-699.